ArchiMob: ein multidialektales Korpus schweizerdeutscher Spontansprache

Scherrer, Yves ; Samardžić, Tanja ; Glaser, Elvira

Abstract: Although Swiss dialects of German are widely used in everyday communication, automatic processing of Swiss German is still a considerable challenge due to the fact that it is mostly a spoken variety and that it is subject to considerable regional variation. This paper presents the ArchiMob corpus, a freely available general-purpose corpus of transcribed spoken Swiss German based on oral history interviews. The corpus is a result of a long design process, intensive manual work and specially adapted computational processing. We first present the modalities of access of the corpus for dialectological, historical and computational research. We then describe how the documents were transcribed, segmented and aligned with the sound source, and summarise a series of experiments that have led to automatically annotated normalisation and part-of-speech tagging layers. Finally, we present several case studies to stimulate the use of the corpus for dialectological research.

DOI: https://doi.org/10.13092/lo.98.5947

ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache*

Yves Scherrer (Helsinki), Tanja Samardžić (Zürich), Elvira Glaser (Zürich)

Abstract

Although Swiss dialects of German are widely used in everyday communication, automatic processing of Swiss German is still a considerable challenge due to the fact that it is mostly a spoken variety and that it is subject to considerable regional variation. This paper presents the ArchiMob corpus, a freely available general-purpose corpus of transcribed spoken Swiss German based on oral history interviews. The corpus is a result of a long design process, intensive manual work and specially adapted computational processing. We first present the modalities of access of the corpus for dialectological, historical and computational research. We then describe how the documents were transcribed, segmented and aligned with the sound source, and summarise a series of experiments that have led to automatically annotated normalisation and part-of-speech tagging layers. Finally, we present several case studies to stimulate the use of the corpus for dialectological research.

1 Einleitung

Obwohl der Dialektgebrauch in der Deutschschweiz zum Alltag gehört, sind digitale Ressourcen für die dialektologische und computerlinguistische Forschung nur begrenzt verfügbar. Traditionell werden die Anwendungsgebiete des Standarddeutschen und der schweizerdeutschen Dialekte nach dem Konzept der „medialen Diglossie“ aufgeteilt (cf. Kolde 1981; Siebenhaar/Wyler 1997), wobei in der schriftlichen Kommunikation (und in einigen institutionalisierten Bereichen der mündlichen Kommunikation) das Standarddeutsche verwendet wird und in der mündlichen Kommunikation die verschiedenen Dialekte. Mit der Entwicklung von computergestützter Kommunikation ist diese traditionelle Aufteilung allerdings aufgeweicht geworden, da schweizerdeutsche Dialekte zunehmend verschriftlicht werden (cf. Siebenhaar 2003). Diese Entwicklungen verlangen und ermöglichen gleichzeitig die automatische Verarbeitung des Schweizerdeutschen, für die Forschung in den Digital Humanities im Allgemeinen und für die korpusgestützte Dialektologie im Speziellen.

* Danksagung: Wir danken unseren zahlreichen Mitwirkenden, die an der Entstehung dieses Korpus beteiligt waren: Noëmi Aepli, Henning Beywl, Christof Bless, Alexandra Bünzli, Matthias Friedli, Anne Göhring, Noemi Graf, Anja Hasse, Gordon Heath, Agnes Kolmer, Mike Lingg, Patrick Mächler, Eva Peters, Beni Ruef, Hanna Ruch, Franziska Schmid, Fatima Stadler, Janine Steiner-Richter, Phillip Ströbel, Simone Ueberwasser, Alexandra Zoller. Dieser Artikel basiert auf einem demnächst in der Zeitschrift *Language Resources and Evaluation* erscheinenden englischsprachigen Beitrag; wir bedanken uns bei Raffaela Zaugg für die Mithilfe bei der Übersetzung ins Deutsche.

Dieser Beitrag stellt das ArchiMob-Korpus vor, ein annotiertes Korpus schweizerdeutscher Spontansprache. Wir nutzen Werkzeuge der natürlichen Sprachverarbeitung, um zusätzliche Annotationsebenen bereitzustellen und zeigen mit einigen Fallbeispielen die Eignung des Korpus für Forschung in den Sprach- und Geisteswissenschaften. Dabei werden gezielt die Herausforderungen der Digitalisierung einer heterogenen Gruppe von Sprachvarietäten ohne schriftliche Tradition angegangen.

Die grössten schweizerdeutschen Ressourcen wurden im Kontext der Dialektologie erstellt und bestehen aus mehr oder weniger isolierten Worttypen, wie im Schweizerdeutschen Wörterbuch *Idiotikon* (Staub et al. 1881ff.) und im Sprachatlas der deutschen Schweiz (Hotzenköcherle et al. 1962–1997; Christen et al. 2013); neuere digitale Ressourcen derselben Art sind u. a. Leemann et al. (2016).

Schweizerdeutsche Textkorpora wurden zum Beispiel im Rahmen von Christen (1998) (transkribierte Interviews) oder Siebenhaar (2003) (Chat-Beiträge) kompiliert. In neuerer Zeit hinzugekommen sind unter anderem ein Korpus von SMS-Nachrichten (Stark et al. 2009–2015), ein Korpus von schriftlichen Dialekttexten (Hollenstein und Aepli 2014, 2015) und ein monodialektales Korpus von transkribierten Interviews (Schönenberger und Haeberli, erscheint). Das ArchiMob-Korpus unterscheidet sich von diesen existierenden Ressourcen dahingehend, dass es alle folgenden Kriterien gleichzeitig erfüllt: es deckt mehrere Dialektgebiete ab, ist mit den Audiodaten aligniert, enthält detaillierte Metadaten zu den Sprechenden, repräsentiert transkribierte gesprochene Sprache, ist vom Inhalt her homogen (historische Erzählungen) und ist nicht zuletzt frei verfügbar.

Dieser Artikel beginnt mit einer Inhaltsdarstellung des ArchiMob-Korpus und seinen Zugriffsmodalitäten (2. Kapitel). Im 3. Kapitel werden die Kodierungs- und Annotationsebenen des Korpus näher beschrieben und unsere Experimente zur Automatisierung der Annotationen zusammengefasst. Schliesslich stellt Kapitel 4 einige Fallstudien vor, die auf der Grundlage des ArchiMob-Korpus verschiedene Aspekte sprachlicher Variation untersuchen und so die Bedeutung der Ressource für die Digital Humanities und die korpusgestützte Dialektologie aufzeigen.

2 Das ArchiMob-Korpus: Von der Zeitzeugenbefragung zur digitalen Forschungsressource

1998 wurde vom Filmemacher Frédéric Gonseth das Projekt Archimob initiiert und der gleichnamige Verein gegründet.[1] Ziel dieser Zusammenarbeit zwischen Historikern und Filmemachern war es, Zeitzeugnisse über persönliche Erfahrungen des Lebens in der Schweiz während des Zweiten Weltkrieges zu sammeln. Dabei entstand eine Sammlung von 555 Interviewaufnahmen, die unter anderem politische Streitigkeiten, verbotene Liebe und das tägliche Leben in der Kriegszeit thematisieren. 300 der 555 Interviews sind auf Schweizerdeutsch geführt worden. Jedes Gespräch wurde mit einer Gewährsperson in halb-direktiver Technik geführt und dauerte in der Regel eine bis zwei Stunden. Die Gewährspersonen

[1] Archimob steht für *archives de la mobilisation* oder *Archive der Mobilmachung*. Wir verwenden die Schreibweise *Archimob* für den Verein und das Datenerfassungsprojekt und *ArchiMob* für das Korpus. Links zu den entsprechenden Webseiten befinden sich im Literaturverzeichnis.

stammen aus allen Regionen der Schweiz und vertreten beide Geschlechter, unterschiedliche soziale Schichten und unterschiedliche politische Ansichten. Die meisten Auskunftspersonen wurden zwischen 1910 und 1930 geboren.

Die Zusammenstellung des ArchiMob-Korpus begann 2004, als vom Verein Archimob eine Sammlung von 52 VHS-Kassetten mit ebensovielen Interviews bezogen werden konnte. Bei der Auswahl der Interviews wurde darauf geachtet, dass die Befragten keinem starken Dialekt- oder Sprachkontakt ausgesetzt waren, was die Aufnahmen für die dialektologische Forschung weniger interessant gemacht hätte. Zunächst war geplant, das Korpus für Untersuchungen in der Dialektsyntax zu verwenden, als alternative Datenquelle zu den Erhebungen des *Syntaktischen Atlas der deutschen Schweiz* (cf. Glaser/Bart 2015). Das ArchiMob-Material wurde zum Beispiel in Studien zur Position des unbestimmten Artikels in adverbial ergänzten Nominalphrasen (cf. Richner-Steiner 2011) und zur Konstruktion von Vergleichssätzen (cf. Friedli 2012) benutzt.

Von den 52 Aufnahmen wurden neun von der weiteren Bearbeitung ausgeschlossen, entweder wegen schlechter Tonqualität oder aufgrund des obengenannten Dialektkontakt-Kriteriums. Die restlichen 43 Aufnahmen wurden dann im MP4-Format digitalisiert.

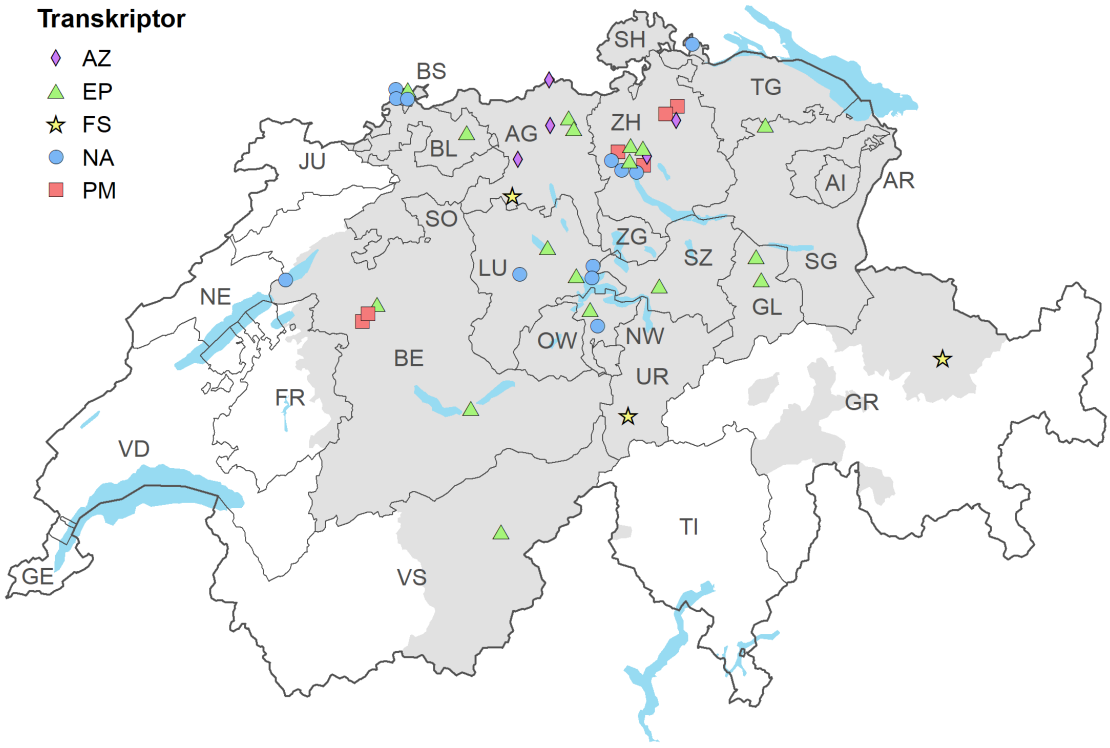


Abbildung 1: Sprecherherkunftsorte der im ArchiMob-Korpus enthaltenen Aufnahmen. Die verschiedenen Symbole entsprechen verschiedenen Transkriptoren (siehe Abschnitt 3.1). Der deutschsprachige Teil der Schweiz ist grau hinterlegt.

Eine erste Version des Korpus mit 34 Aufnahmen und durchschnittlich 15‘540 Tokens pro Aufnahme wurde 2016 veröffentlicht (cf. Samardžić et al. 2016). Die zweite Version, die in diesem Artikel beschrieben wird, enthält alle 43 ausgewählten Aufnahmen. Abbildung 1 zeigt

die räumliche Verteilung der im Korpus enthaltenen Aufnahmen nach Herkunft der Gewährspersonen.

Die ausgewählten Aufnahmen wurden dann so transkribiert und verarbeitet, dass sie nach verschiedenen Aspekten durchsucht werden können. Die nachfolgend näher beschriebenen Verarbeitungsschritte umfassen die Normalisierung und das Part-of-speech-Tagging.

Um den unterschiedlichen Bedürfnissen der Nutzer gerecht zu werden, stellen wir das Korpus auf zwei Arten zur Verfügung: als Online-Version über Korpus-Abfragesysteme und als XML-Download. Aktuell kann über die Projektwebseite (siehe Literaturverzeichnis) auf das Korpus zugegriffen werden. Die Audiodateien sind auf Anfrage erhältlich.[2]

2.1 Online-Zugriff mit Korpus-Abfragesystemen

Wir haben uns für die parallele Nutzung von zwei Korpus-Abfragesystemen entschieden: Sketch Engine (cf. Kilgarriff et al. 2014) und ANNIS (cf. Krause/Zeldes 2016).[3]

Ein Suchergebnis-Beispiel mit Sketch Engine ist in Abbildung 2 dargestellt. Das System zeigt nicht nur Textpassagen mit der exakten Übereinstimmung des Suchwortes *gält* ‚Geld' an, sondern auch andere dialektale Varianten wie *gäld* oder *gäut*. Eine solche flexible Suchmöglichkeit ist in der einfachen Suchoption voreingestellt. Um verschiedene Varianten desselben Wortes miteinander in Beziehung zu setzen, verwenden wir eine normalisierte Schreibweise, die in Abbildung 2 als grauer Index des Suchwortes dargestellt ist. Diese normalisierte Schreibweise entspricht in den meisten, aber nicht in allen Fällen der standarddeutschen Schreibweise (siehe Abschnitt 3.2).

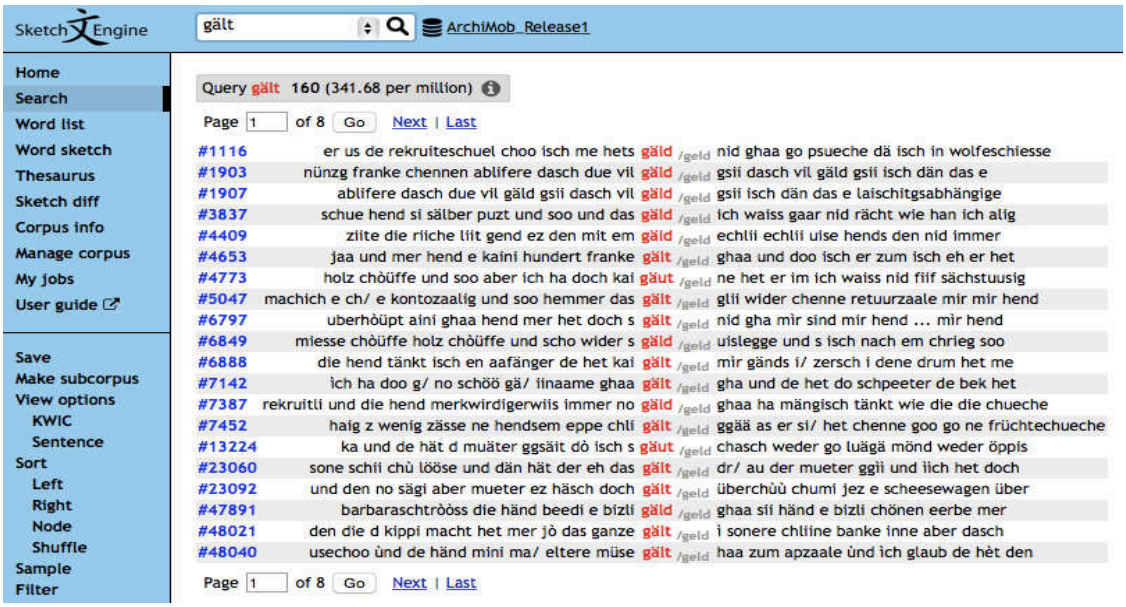


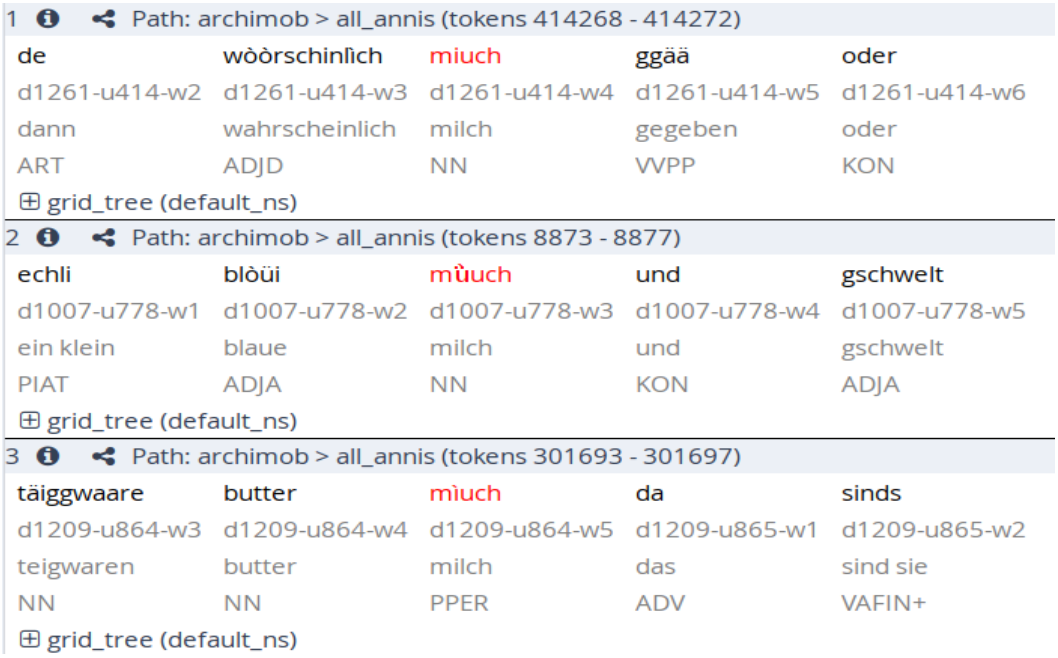
Abbildung 2: Beispiel eines Abfrageergebnisses mit Sketch Engine für das Wort *gält* ‚Geld'.

[2] Das XML-Archiv enthält ein paar Beispiel-Audiodateien zur Evaluation.

[3] Die Zugriffsmodalitäten auf diese Korpus-Abfragesysteme sind auf der Projektwebseite genauer beschrieben.

Es kann und soll nicht davon ausgegangen werden, dass die Benutzer die genaue Normalisierung eines Wortes kennen oder gar die projektspezifischen Normalisierungsrichtlinien erlernen. Um es den Benutzern dennoch zu ermöglichen, das Korpus effizient zu durchsuchen, wurde vom Sketch Engine-Team speziell für unser Projekt eine neue Funktion implementiert. Diese Funktion ermöglicht es den Benutzern, die Abfrage in jeglicher Schreibweise einzugeben, die ihnen plausibel erscheint. Wenn diese Schreibweise mindestens einmal im Korpus vorkommt, werden die Beispiele des abgefragten Begriffs mit allen anderen Schreibweisen verlinkt und angezeigt. Diese flexible, dialektunabhängige Suche unterscheidet sich von den üblichen Suchmethoden: da Korpusabfragesysteme in erster Linie für die Arbeit mit Texten in Standardsprachen konzipiert wurden, erwarten sie, dass die Benutzer die genaue Schreibweise für die Abfrage kennen oder reguläre Ausdrücke verwendet, um die Funktionalität der flexiblen Suche zu simulieren. Unsere Lösung ermöglicht die Suche nach Quellen mit inkonsistenter Schreibweise auf intuitive und benutzerfreundliche Weise und macht sie einem breiteren Publikum zugänglich. Diese Funktion ist daher unter Umständen nicht nur für das Schweizerdeutsche, sondern für alle nicht standardisierten Sprachen und Varietäten von Nutzen.

Zusätzlich zur neuen flexiblen Suchfunktion können Nutzer von Sketch Engine die Standardfunktionalität des Systems verwenden, um erweiterte Abfragen mit Hilfe der Korpus-Abfragesprache auszuführen, Konkordanzen zu bearbeiten und verschiedene Statistiken zu berechnen (z. B. signifikante Kollokationen).



1 Path: archimob > all_annis (tokens 414268 - 414272)

de	wòòrschinlïch	miuch	ggää	oder
d1261-u414-w2	d1261-u414-w3	d1261-u414-w4	d1261-u414-w5	d1261-u414-w6
dann	wahrscheinlich	milch	gegeben	oder
ART	ADJD	NN	VVPP	KON

⊞ grid_tree (default_ns)

2 Path: archimob > all_annis (tokens 8873 - 8877)

echli	blòüi	mùuch	und	gschwelt
d1007-u778-w1	d1007-u778-w2	d1007-u778-w3	d1007-u778-w4	d1007-u778-w5
ein klein	blaue	milch	und	gschwelt
PIAT	ADJA	NN	KON	ADJA

⊞ grid_tree (default_ns)

3 Path: archimob > all_annis (tokens 301693 - 301697)

täiggwaare	butter	mìuch	da	sinds
d1209-u864-w3	d1209-u864-w4	d1209-u864-w5	d1209-u865-w1	d1209-u865-w2
teigwaren	butter	milch	das	sind sie
NN	NN	PPER	ADV	VAFIN+

⊞ grid_tree (default_ns)

Abbildung 3: Beispiel eines Abfrageergebnisses mit ANNIS: die Suche nach der normalisierten Form *milch* ergibt Treffer mit den dialektalen Varianten *miuch* und *müuch*. Es ist zu beachten, dass dieses Beispiel einem Satz entstammt, der automatisch normalisiert und annotiert wurde (siehe Abschnitt 3.2.2 und 3.3.2) und deshalb einige Fehler enthält. In Zeile 1 sollte *de* als Adverb statt Artikel getaggt werden und in Zeile 3 *mìuch* als Nomen statt Personalpronomen. Überdies illustriert die Normalisierung *blaue* in Zeile 2, dass unser Normalisierungsansatz (siehe Abschnitt 3.2) nicht mit einer Übersetzung ins Standarddeutsche gleichgesetzt werden darf – eine korrekte Übersetzung wäre nämlich *geschlagen*.

Um dem Bedürfnis einiger potentieller Nutzer des ArchiMob-Korpus nach einer detaillierteren Visualisierung der Suchergebnisse gerecht zu werden, verwenden wir alternativ das Korpus-Abfragesystem ANNIS. Ein Beispiel für ein ANNIS-Abfrageergebnis ist in Abbildung 3 dargestellt. Man sieht, dass das System alle derzeit im Korpus verfügbaren Informationen gleichzeitig anzeigt. Unter jedem transkribierten Wort sehen wir seine Korpus-ID, die normalisierte Schreibweise und das Part-of-speech-Tag. Die Anzahl der angezeigten Treffer bleibt dabei jedoch gering, da eine solche Detailansicht den Bildschirm schnell ausfüllt.

Wir haben die flexible Suchmöglichkeit in ANNIS nicht implementiert, da dieses System für fortgeschrittene Nutzer gedacht ist, die sich für linguistische Details interessieren. Dies zeigt sich nicht nur in der detaillierten Darstellung, sondern auch in der etwas schwierigeren Handhabung der Abfragesprache.

Ein wichtiger Unterschied zwischen den beiden Korpus-Abfragesystemen ist, dass Sketch Engine-Nutzer von ausserhalb der Europäischen Union ein kostenpflichtiges Konto für den ArchiMob-Zugriff benötigen, während persönliche Konten auf ANNIS kostenlos sind.[4]

2.2 XML-Download

Neben der Online-Suche bieten wir ein XML-Archiv zum Herunterladen an. Das XML-Format folgt grösstenteils den Empfehlungen der *Text Encoding Initiative* (TEI). Diese XML-Dateien liegen auch den Korpus-Abfragesystemen zu Grunde. Die Daten werden in drei verschiedenen Dateitypen gespeichert:

- **Inhaltsdateien** beinhalten den Text der Transkriptionen.
- **Mediendateien** enthalten Informationen zur Alignierung des transkribierten Texts mit den entsprechenden Audiodateien.
- **Sprecherdateien** enthalten soziodemografische Informationen über die Gewährspersonen (Region/Dialekt, Alter, Geschlecht, Beruf) und Informationen über deren Rolle im Gespräch (Interviewer/in, Befragte/r).

Die Inhaltsdateien sind in Äusserungen (XML-Element ‘u’) unterteilt. Die Verweise auf den Sprecher/die Sprecherin und die Mediendatei werden in jeder Äusserung folgendermassen als Attribute angegeben:

- <u start="mediapointers#d1007-T176" xml:id="d1007-u88" who="persondb#EJos1007">

Äusserungen bestehen aus Wörtern (XML-Element ‚w‘). Die normalisierte Form und das Part-of-speech-Tag werden als Attribute kodiert:

- <w normalised="einest" POS="ADV" xml:id="..."> ainisch</w>

Die Äusserungen enthalten auch Pausen (gefüllt oder leer), Wiederholungen und unklare (oder nicht transkribierbare) Stellen. Pausen werden nicht als Wörter gezählt, sondern mit einem anderen XML-Element (‘pause’) versehen:

- <w normalised="wo" POS="KOUS" xml:id="...">won</w>

[4] Eine frei zugängliche Open-Source-Version von Sketch Engine ist unter dem Namen NoSketch verfügbar (cf. Rychly 2007). Zurzeit verwenden wir diese Version nicht.

- <w normalised="ich" POS="PPER" xml:id="...">ìch</w>
- <pause vocal="eh" />
- <w normalised="ja" POS="ITJ" xml:id="...">ja</w>

Unklare Sprache wird in ein spezifisches XML-Element eingefasst, das sich über mehrere Wörter erstrecken kann. Bei einer Wiederholung wird das betreffende Wort nur einmal als Wort annotiert; die wiederholten Teile werden als Löschung (XML-Element 'del') annotiert:

- <del type="truncation" xml:id="...">hundertvierz/</del>
- <del type="truncation" xml:id="...">hundertvier/</del>
- <w normalised="hundertfünfundvierzig" tag="NN" xml:id="...">hundertfiifevierzgi</w>

Die Medien- und Sprecherdateien sind einfache XML-Dokumente, die Listen von Zeit- und Sprecher-IDs und deren Informationen enthalten.

3 Korpus-Kodierung und Annotation

Die Umwandlung von Aufnahmen gesprochener Sprache in eine allgemein zugängliche Forschungsressource erfordert eine umfangreiche Bearbeitung. In diesem Abschnitt beschreiben wir die Kodierungs- und Annotationsschritte, die bei der Erstellung des ArchiMob-Korpus angewandt wurden, und gehen auf die spezifischen Herausforderungen des Schweizerdeutschen ein.

3.1 Transkription und Alignierung der Ton- und Textdaten

Die 43 für das Korpus ausgewählten Dokumente wurden von fünf Transkribierenden in vier Phasen transkribiert. Die Transkriptionsmodalitäten und -phasen waren nicht Teil eines übergeordneten Plans, sondern das Ergebnis verschiedener Umstände, unter denen die Arbeit am Korpus stattfand. Tabelle 1 gibt einen Überblick über die Laufzeit des Annotationsprozesses. Die Tabelle im Anhang enthält weitere Kenndaten der Dokumente.

Phase/Jahre	Transkriptor/Dokument-IDs		Transkriptionswerkzeug
1 2006–2012	EP	1007, 1048, 1063, 1073, 1075, 1142, 1143, 1147, 1170, 1195, 1198, 1207, 1209, 1212, 1261, 1270	Nisus Writer
2 2012–2014	PM	1082, 1083, 1087, 1121, 1215, 1225, 1244	FOLKER
3 2015	NA	1008, 1055, 1138, 1188, 1189, 1205	EXMARaLDA
	AZ	1228, 1248, 1259, 1295, 1300	
4 2016–2017	NA	1044, 1053, 1203, 1224, 1235, 1263	EXMARaLDA
	FS	1163, 1240, 1255	

Tabelle 1: Überblick über die vier Transkriptionsphasen.

Der transkribierte Text wurde in Äusserungen unterteilt, wobei wir unter einer Äusserung eine durch die Prosodie (v. a. Pausen) definierte Transkriptionseinheit mit einer durchschnittlichen Länge von ca. 4–8 Sekunden verstehen. Äusserungen erstrecken sich in der Regel über eine oder mehrere syntaktische Phrasen. In Korpora gesprochener Sprache ist es üblich, benachbarte Äusserungen desselben Sprechers zu Turns zu gruppieren. Wir haben die Grenzen zwischen den Turns nicht explizit gekennzeichnet, aber da jede Äusserung mit einer Sprecher-ID versehen ist, können Turn-Grenzen leicht aufgrund des Wechsels der Sprecher-ID erkannt werden.

Die mit der Tonquelle alignierten Transkriptionseinheiten wurden von den Transkribierenden manuell gebildet. Diese Alignierung konnte direkt in spezialisierten Annotationswerkzeugen wie FOLKER und EXMARaLDA (cf. Schmidt 2012, für beide Werkzeuge) vorgenommen werden. Da in Phase 1 kein spezialisiertes Werkzeug eingesetzt wurde, mussten die 16 in dieser Phase erstellten Dokumente nachträglich mit der Tonquelle aligniert werden. Dazu alignierten wir die Transkriptionen zunächst automatisch – mit Hilfe des Programms WebMAUS (cf. Kisler et al. 2012) – mit der Tonquelle. Um eine der manuellen Alignierung vergleichbare Granularität und Qualität zu erhalten, fügten wir die WebMAUS-Alignments automatisch zu grösseren Einheiten zusammen und importierten sie zur manuellen Korrektur in EXMARaLDA. Bei etwa einem Drittel der Transkriptionen funktionierte die WebMAUS-Alignierung nicht gut genug. In diesen Fällen erstellten wir zunächst eine Annäherung der Transkriptionseinheiten basierend auf den im Transkript kodierten Pausen und importierten sie dann zur manuellen Korrektur in EXMARaLDA.

Für die Transkription der Texte verwendeten wir die von Dieth (1986) vorgeschlagene *Schwyzertütschi Dialäktschrift*, wie es in der aktuellen dialektologischen Forschung üblich ist. Die Transkription soll die wichtigsten phonetischen Eigenschaften der Varietät anzeigen, jedoch auch für alle, die mit der standarddeutschen Schreibweise vertraut sind, lesbar sein (cf. Dieth 1986: 10). Die Funktion des Grapheminventars in Dieths Schrift hängt vom Dialekt und dessen phonetischen Eigenschaften ab. Das Graphem ‹e› steht beispielsweise für die unterschiedlichen Vokaleigenschaften [e], [ɛ] oder [ə], je nach Dialekt, Betonung der Silbe und – in erheblichem Masse – dialektalem Hintergrund der Transkriptorin oder des Transkriptors.

Das ursprünglich phonemische System von Dieth kann auf unterschiedliche Weise implementiert werden, je nachdem, wie differenziert die phonetischen Eigenschaften ausgedrückt werden sollen. In unserem Projekt hat sich die Anwendungspraxis von Dieths System über die Transkriptionsphasen hinweg geändert, so dass in der ersten Phase mehr Unterscheidungen bezüglich Vokalqualität und -quantität gemacht wurden als in den Folgephasen (z. B. Phase 1: *èèr* vs. Phase 3: *er* ‚er').[5]

Eine Durchsicht der Transkriptionen ergab, dass diese Änderungen in den Leitlinien nicht die einzige Quelle für Abweichungen waren. Die verschiedenen Transkribierenden setzten die Leitlinien in unterschiedlicher Weise um, nicht nur in Bezug auf die Vokaleigenschaften, sondern auch in Bezug auf die Wortsegmentierung und andere Aspekte. Diese Inkonsistenz ist

[5] Die Projektwebseite enthält weiterführende Informationen zu unseren Transkriptionsleitlinien.

einer der Gründe, weshalb wir eine zusätzliche Annotationsebene von normalisierten Wortformen einführen (siehe unten).

Alle in der aktuellen Version des Korpus enthaltenen Transkriptionen wurden manuell mit den in Tabelle 1 aufgeführten Transkriptionswerkzeugen erstellt. Diese Transkriptionen bilden jedoch ein erstes Trainingsset, um in Zukunft weitere Dialekttexte ausserhalb des ArchiMob-Korpus mit automatischen Spracherkennungssystemen zu verarbeiten. In Bezug auf das vorliegende Korpus könnten damit zum Beispiel Transkriptoreneffekte ohne zusätzliche manuelle Transkriptionsarbeit erkannt und automatisch reduziert werden. Erste Experimente und Resultate werden in Scherrer et al. (erscheint) vorgestellt. In einer subjektiven Beurteilung durch unsere Transkribierenden werden die aktuellen Resultate des Systems noch als unzureichend beurteilt, aber wir werden weiter daran arbeiten, das Spracherkennungssystem durch die Einführung von zusätzlichen Trainingsdaten und neuen Lernmethoden zu verbessern.

3.2 Normalisierung

Variation in der Verschriftlichung des Schweizerdeutschen entsteht auf zwei Ebenen. Erstens bewirkt die dialektale Variation, dass lexikalische Einheiten in verschiedenen Regionen unterschiedlich ausgesprochen und damit auch unterschiedlich geschrieben werden. Zweitens werden auch lexikalische Einheiten, die als phonetisch invariant (innerhalb einer Region) betrachtet werden können, nicht immer gleich verschriftlicht, entweder aufgrund gelegentlicher sprecherinterner Abweichungen oder, wie oben erwähnt, aufgrund von Transkriptoreneffekten. Um eine lexikalische Identität aller Schreibvarianten zu schaffen, die intuitiv ‚das gleiche Wort' darstellen – zum Beispiel um eine flexible Suche zu ermöglichen –, müssen sie zu einer einzigen Form hin normalisiert werden.

Original	min	maa	het	immer	Gsaait
Varianten im gleichen Dokument	mi mii miin		*hat*		Gsait
Varianten in anderen Dokumenten	mine mìì *mäin* *main* *mein*	ma *man* *mann*	hed hèd hèt hät hätt	ime imer emmer imme immers	gsäit gsääit gseit gseid ggsait
Normalisierung	mein	mann	hat	immer	Gesagt

Tabelle 2: Ein Abschnitt eines transkribierten Textes (Original) mit den entsprechenden Varianten, die im gleichen und in anderen Dokumenten vorkommen. Die kursiv gesetzten Varianten stellen Transkriptionen standarddeutsch ausgesprochener Wörter in Code-Switching- und Zitats-Kontexten dar (mein Gott, Mein Kampf, Not am Mann, …),. Die unterste Zeile zeigt die gemeinsame Normalisierung aller Varianten.

Tabelle 2 zeigt die Bandbreite der möglichen Variation in einer beliebig gewählten Äusserung aus unserem Korpus. Die Tabelle zeigt alle Varianten der ausgewählten Wörter im gleichen Dokument, i. e. von der gleichen Gewährsperson produziert und vom gleichen Experten transkribiert. Zusätzlich finden sich weitere Varianten in anderen Dokumenten, die Beispiele anderer Varietäten enthalten. Zu den aufgeführten Varianten gehören Fälle von dialektaler Variation (z. B. *gsait*, *gsäit*), Varianten aufgrund geänderter Transkriptions-Leitlinien (z. B. *hed*, *hèd*) und Varianten aufgrund von Code-Switching (z. B. *mäin, main, mann, hat*).

Im Beispiel von Tabelle 2 entsprechen alle normalisierten Formen standarddeutschen Formen. Tatsächlich wird für die Normalisierung immer dann eine standarddeutsche Form verwendet, wenn sie der dialektalen Wortform in Bedeutung und Etymologie entspricht. Wir verstehen Normalisierung allerdings nicht als Übersetzung ins Standarddeutsche: eine echte Übersetzung würde erhebliche syntaktische Transformationen und lexikalischen Austausch erfordern, wohingegen die Normalisierung eher einer Wort-für-Wort-Annotation der lexikalischen Identität entspricht. Die exakte Wahl der normalisierten Formen kann dabei als willkürlich angesehen werden. Einige unserer Entscheidungen für die Normalisierungsebene bedürfen weiterer Erklärungen:

- Schweizerdeutsche Wortformen, die keine etymologisch verwandten standarddeutschen Pendants haben, werden durch eine rekonstruierte gemeinsame schweizerdeutsche Form normalisiert. Zum Beispiel wird *öpper* als *etwer* anstelle des semantischen standarddeutschen Äquivalents *jemand* normalisiert, *töff* als *töff* anstelle des standarddeutschen Wortes *Motorrad*, *gheie* ‚fallen' als *geheien*. Ebenso wird beispielsweise schweizerdeutsches *vorig* ‚übrig' als *vorig* normalisiert, obwohl es im Standarddeutschen ‚vorhergehend' bedeutet.
- Standarddeutsche Konventionen bezüglich Wortgrenzen sind auf das Schweizerdeutsche oft nicht anwendbar, da Artikel und Pronomen häufig klitisiert werden. Das hat zur Folge, dass Transkribierende Formen vorschlagen, die mehreren standarddeutschen Tokens entsprechen. In solchen Fällen erlauben wir pro transkribiertes Token mehrere Tokens auf der normalisierten Seite. Beispielsweise wird *hettemers* als *hatten wir es* normalisiert und *bimene* als *bei einem*.
- Manchmal hat die Normalisierung den willkommenen Nebeneffekt, homophone (bzw. homographe) Dialektformen zu disambiguieren. Beispielsweise wird *de* je nach Kontext als *der* (bestimmter Artikel) oder als *dann* (Temporaladverb) normalisiert.
- In anderen Fällen umfasst eine normalisierte Form aufgrund von morphosyntaktischem Synkretismus formal unterschiedliche Dialektformen. Beispielsweise wird das erste normalisierte Wort aus Tabelle 2, *mein*, nicht nur auf die maskulinen Dialektformen *min, miin* angewendet, sondern auch auf die neutralen Formen *mis*, *miis*.

Ein wichtiges Merkmal unseres Ansatzes ist, dass wir die Normalisierung als eine versteckte Annotationsebene betrachten, die in erster Linie für die automatische Verarbeitung verwendet wird. Wie oben erwähnt, wird von den Nutzern nicht erwartet, dass sie ihre Abfragen in der Normalisierungssprache formulieren.

3.2.1 Manuelle Normalisierung

Unser Normalisierungsansatz wurde in detaillierten Richtlinien zusammengefasst, die wir dann auf die manuelle Normalisierung von zunächst sechs Dokumenten aus der Transkriptionsphase 1 (Dokument-IDs 1007, 1048, 1063, 1143, 1198, 1270, siehe Anhang für Details) von drei erfahrenen Annotierenden anwendeten. Für diese Aufgabe benutzten wir Annotationswerkzeuge, die es den Annotierenden ermöglichten, schnell frühere Normalisierungen für bereits normalisierte Wörter nachzuschlagen. Wir verwendeten zuerst VARD 2 (cf. Baron/Rayson 2008), wechselten aber später auf das passendere SGT-Tool (cf. Ruef/Ueberwasser 2013).

3.2.2 Automatisierung der Normalisierung mittels maschineller Übersetzung

Um auch die verbleibenden Dokumente mit Normalisierungen versehen zu können, verwendeten wir Methoden der zeichenbasierten statistischen maschinellen Übersetzung (CSMT, für *character-level statistical machine translation*). Dabei dienten uns die manuell normalisierten Dokumente als Trainingsmaterial des Übersetzungssystems.

CSMT wurde ursprünglich für die Übersetzung zwischen eng verwandten Sprachen vorgeschlagen (cf. Vilar et al. 2007; Tiedemann 2009). Sie benötigt weniger Trainingsdaten als die herkömmliche wortbasierte statistische maschinelle Übersetzung, ist aber auf Anwendungen beschränkt, bei denen reguläre Veränderungen auf Zeichenebene stattfinden. In jüngster Zeit wurde CSMT erfolgreich zur Normalisierung bei computervermittelter Kommunikation (cf. z. B. De Clercq et al. 2013; Ljubešić et al. 2014; Ljubešić et al. 2016) und bei historischen Texten (cf. Pettersson et al. 2013, 2014; Scherrer/Erjavec 2016; Ljubešić et al. 2016; Tjong Kim Sang et al. 2017) eingesetzt. Unsere verschiedenen Experimente zur automatischen Normalisierung der ArchiMob-Daten werden detailliert in Samardžić et al. (2015), Samardžić et al. (2016) und Scherrer/Ljubešić (2016) beschrieben.

Bei der manuellen Überprüfung der ersten Ergebnisse stellte sich heraus, dass die anfänglichen Normalisierungsrichtlinien nicht klar genug waren, um eine konsistente Annotation zu gewährleisten. Beispielsweise wurde die eindeutige schweizerdeutsche Form *dra* manchmal als *dran* und manchmal als *daran* normalisiert; beide Normalisierungen sind korrekte standarddeutsche Wörter. Auch die schweizerdeutsche Form *gschaffet* wurde manchmal auf das standarddeutsche semantische Äquivalent *gearbeitet* und manchmal auf sein etymologisches Äquivalent *geschafft* normalisiert. Nach einer Überarbeitung der entsprechenden Leitlinien (in den konkreten Fällen bevorzugten wir die längere Form *daran* sowie das etymologische Äquivalent *geschafft*) und der Adaptierung der bereits annotierten Daten konnten wir die Normalisierungsgenauigkeit von 77.28% auf 84.13% erhöhen. Eine Weiterentwicklung des Normalisierungswerkzeugs brachte eine weitere Erhöhung der Genauigkeit auf 90.46% mit sich.

Honnet et al. (2018) und Lusetti et al. (2018) zeigen, dass in leicht anders gelagerten Normalisierungsaufgaben die besten Ergebnisse mit neuronalen Methoden erzielt werden können. Wir beabsichtigen, in Zukunft neuronale Methoden auch auf unsere Definition der Normalisierung anzuwenden.

3.3 Part-of-speech-Tagging

Die syntaktische Annotation in Form von Part-of-speech-Tags ist wichtig, um abstraktere Korpusabfragen zu ermöglichen, die Wortklassen und deren Kombinationen betreffen. Part-of-speech-Tagging ist in der automatischen Sprachverarbeitung gut erforscht, und es stehen zahlreiche Werkzeuge zur Verfügung. Diese werden jedoch in der Regel für geschriebene, standardisierte Sprachen entwickelt und getestet, während wir sie auf eine gesprochene, nicht standardisierte Varietät anwenden können müssen. Wir stützen uns dabei auf Hollenstein/Aepli (2014), die das weit verbreitete Stuttgart-Tübingen-Tagset (STTS) (cf. Thielen et al. 1999) auf (geschriebenes) Schweizerdeutsch angepasst haben. Die Anpassung des Tagsets betrifft folgende in den Dialekten beobachtete Phänomene:

- Das neue Tag PTKINF wird für die Infinitivpartikeln *go, cho, la, afa* eingeführt. Diese Partikeln werden verwendet, wenn die jeweiligen Vollverben (gehen, kommen, lassen, beginnen) einen Infinitivsatz subkategorisieren. Da dieses Phänomen im Standarddeutschen nicht existiert, ist die Einführung eines neuen Tags erforderlich.
- Das Tag APPRART, das im Standarddeutschen für Präposition + bestimmter Artikel verwendet wird, wird auf Präposition + unbestimmter Artikel erweitert, wie in *bimene* ‚bei einem'.
- Die Bezeichnungen VAFIN+ und VMFIN+ gelten für Verbformen mit Enklitika. Letztere sind in der Regel Pronomen, z. B. *häts* ‚hat es', *hettemers* ‚hätten wir es'. Konjunktionen mit Enklitika werden mit KOUS+ bezeichnet, z. B. *wemmer* ‚wenn wir'.
- Das Infinitivpartikel *zu*, im Schweizerdeutschen reduziert auf *z*, wird zum Infinitiv hinzugefügt und mit dem Tag PTKZU+ versehen: *zflüge* ‚zu fliegen'.
- Adverbien mit Enklitika (Artikel oder andere Adverbien) erhalten das Tag ADV+: *sones* ‚so ein'.

3.3.1 Manuelle Annotation

Wie bei der Normalisierung wird auch hier ein kleiner Teil des ArchiMob-Korpus von Hand mit Part-of-speech-Tags annotiert. Wir nehmen dazu jeweils 10% aus jeder der sechs Dateien, die zuvor manuell normalisiert wurden (Dokument-IDs 1007, 1048, 1063, 1143, 1198, 1270) und annotieren sie mit den STTS+-Tags. Dieses Teilkorpus besteht aus 10'169 Tokens in 1'742 Äusserungen.

3.3.2 Automatisierung der Tag-Annotation

Die Notwendigkeit der manuellen Annotation ist für das Tagging eine etwas andere als für die Normalisierung. Während wir für die automatische Normalisierung sowohl Trainings- und Testbeispiele manuell erstellen mussten, werden für das Tagging nur Testbeispiele benötigt, da wir auf anderweitige Trainingsdaten wie *NOAH's Corpus of Swiss German Dialects* (cf. Hollenstein/Aepli 2014, 2015) oder *TüBa-D/S* (cf. Hinrichs et al. 2000) zurückgreifen können. Erste Experimente (cf. Samardžić et al. 2016) ergaben, dass *NOAH's Corpus* trotz seiner geringeren Grösse die bessere Datenquelle für einen Part-of-speech-Tagger darstellt als *TüBa-D/S*.

Ausgehend von diesem Basis-Tagger führten wir mehrere Korrekturrunden durch. In jeder Korrekturrunde wurde ein ArchiMob-Dokument mit dem Tagger annotiert, die Annotationen manuell korrigiert, das korrigierte Dokument den Trainingsdaten hinzugefügt und ein neuer Tagger damit trainiert. Dieser neue Tagger wurde dann in der folgenden Korrekturrunde verwendet. Nach vier Korrekturrunden konnten wir dabei die Tag-Genauigkeit von 77.18% auf 92.51% steigern. Mit diesem Tagger wurden alle Dokumente des veröffentlichten ArchiMob-Korpus annotiert.

Um das automatische Tagging weiter zu verbessern, versuchten wir, das Trainingsset mit einer aktiven Lernumgebung für manuelle Annotation zu optimieren. Bei diesem Ansatz wurde der Tagger der letzten Korrekturrunde auf alle nicht manuell annotierten oder korrigierten Äusserungen angewendet. Dabei wurden diejenigen Wörter ausgewählt, bei denen der Tagger am unsichersten über seinen eigenen Output war, und für die manuelle Korrektur bereitgestellt. Waren diese Beispiele einmal korrigiert, wurden sie aus dem nicht annotierten Bereich in das Trainingsset verschoben und ein neuer Tagger damit trainiert. Das Verfahren wurde so lange wiederholt, wie es zu sichtbaren Verbesserungen der Annotationsqualität führte. Details zu diesem Verfahren sind in Scherrer et al. (erscheint) beschrieben.

4 Das ArchiMob-Korpus als Grundlage für Sprachvariationsforschung

Das ArchiMob-Korpus ist nicht nur ein interessantes Untersuchungsobjekt für die Computerlinguistik, sondern kann auch als wertvolle Ressource für die dialektologische Forschung dienen, wie dies seit Beginn des Projekts vorgesehen ist, sowohl für qualitative als auch quantitative Forschungsansätze. In diesem Abschnitt stellen wir einige quantitative Fallstudien vor, um das Potenzial des ArchiMob-Korpus zu veranschaulichen. In Abschnitt 4.1 untersuchen wir, inwieweit dialektale Variation allein durch die Betrachtung der Transkriptionen erfasst werden kann. Abschnitt 4.2 befasst sich mit ähnlichen Fragen zur linguistischen Variation, versucht aber, diese unter Berücksichtigung der Normalisierungen zu beantworten. Abschnitt 4.3 veranschaulicht, wie die Annotationen verwendet werden können, um den Inhalt der Texte zu untersuchen.

In allen Fallstudien verwenden wir nur die Äusserungen der Gewährspersonen, nicht diejenigen der Interviewer. So hoffen wir, dass das Datenmaterial den Dialekt der Gewährsperson so gut wie möglich abbildet.

4.1 Bestimmung von dialektalen Variationsmustern in Transkriptionen

Die Transkriptionen des ArchiMob-Korpus liefern eine interessante Datenquelle zur Bestimmung von dialektalen Variationsmustern. Hier werden zwei Arbeitsschritte diskutiert, nämlich den dialektalen Ursprung einer Äusserung zu identifizieren (Dialekterkennung, Abschnitt 4.1.1) und die Dokumente nach ihrer sprachlichen Ähnlichkeit einzuordnen (Dialektklassifikation, Abschnitt 4.1.2).

4.1.1 Dialekterkennung

Spracherkennung ist in der natürlichen Sprachverarbeitung wichtig, und während relativ einfache Methoden bei entsprechend verschiedenartigen Sprachen gut funktionieren, ist die Spracherkennung bei eng verwandten Sprachen immer noch eine Herausforderung (z. B. Zampieri et

al. 2014). Die Herkunftserkennung von Dialekttexten kann dabei als Spezialfall von Spracherkennung bei eng verwandten Sprachen angesehen werden. In diesem Sinne wurden Daten aus dem ArchiMob-Korpus für eine Dialekterkennungsaufgabe im Rahmen der VarDial-Workshops 2017 und 2018 verwendet (cf. Zampieri et al. 2017, 2018).

Die Dokumente aus dem ArchiMob-Korpus wurden manuell in Dialektgebiete gruppiert, wobei vier Bereiche mit ausreichenden Textmengen identifiziert wurden, die sich genug stark voneinander unterscheiden lassen: Zürich (ZH), Basel (BS), Bern (BE) und Luzern (LU). Für jedes Dialektgebiet wurden Äusserungen aus mindestens drei Dokumenten als Trainingsdaten und Äusserungen aus mindestens einem anderen Dokument zum Testen der Systeme ausgewählt (siehe Abbildung 4). Das Ziel war es, das Dialektgebiet einer Äusserung rein auf der Grundlage der in dieser Äusserung verfügbaren linguistischen Merkmale zu bestimmen. Tabelle 3 zeigt einige Beispiele (Details cf. Zampieri et al. 2017, Tabelle (5), sowie Zampieri et al. 2018, Tabelle (4)).

Insgesamt zehn Forschungsgruppen nahmen 2017 mit ihren eigens entwickelten Systemen an der Dialekterkennungsaufgabe teil. Im Durchschnitt erreichten die Systeme eine Genauigkeit von ca. 70% und eine Trefferquote von ca. 70% für alle vier Dialektgebiete, mit einer Ausnahme: Die Trefferquote beim LU-Dialekt war konstant niedrig und lag bei etwa 30%. Grund dafür war die Tatsache, dass LU das einzige Dialektgebiet war, für das keine Äusserungen der Testtranskriptorin (NA) in das Trainingsset aufgenommen worden waren (siehe Abbildung 4). Diese Hypothese wurde dadurch gestützt, dass LU am häufigsten mit BS verwechselt wurde (welches Trainingsdaten von NA enthält, aber dialektologisch eher weit von LU entfernt ist), und dass die Teilnehmenden bei ihren Kreuzvalidierungsexperimenten an den Trainingsdaten keine so geringe Trefferquote festgestellt hatten. Für 2018 wurden zusätzliche Texte in die Datensätze aufgenommen, wobei auf eine gleichmässigere Verteilung der Transkribierenden geachtet wurde.

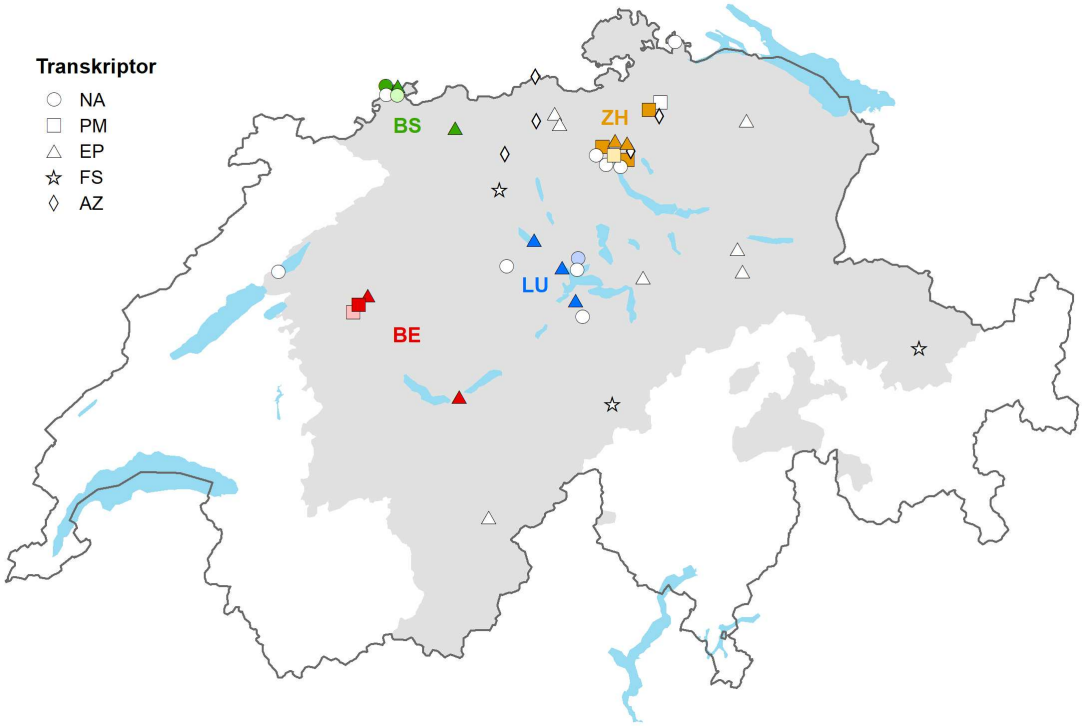


Abbildung 4: ArchiMob-Dokumente, die für die VarDial 2017-Dialekterkennungsaufgabe verwendet wurden. Jedes Symbol stellt ein Dokument dar, mit unterschiedlichen Symbolformen für verschiedene Transkribierende und unterschiedlichen Farben für die vier Dialektzonen (Dokumente mit weissen Symbolen werden in der Dialekterkennungsaufgabe nicht verwendet). Dunkel gefärbte Dokumente dienen als Trainingsdaten, hell gefärbte Dokumente als Testdaten.

u simer geng a d landi öppe zwöi drüümau	BE
unsere suntigschbaziirgang hètis	BS
da häts dänn amel eson en linsebrei ggää	ZH
de vatter hets natüürli glii gmèrkt ùnd dä ìsch ainisch choo ùnd nòchhär nümme	LU
daa bi wind und wätter im schnee	LU
Aaber	BS

Tabelle 3: Beispiele für Äusserungen, die zur Dialekterkennung verwendet werden, mit den korrekten Dialekt-Labels.

Die Werte für die restlichen Dialekte (etwa 70% Genauigkeit und Trefferquote) sind wahrscheinlich nicht wesentlich vom menschlichen Leistungsvermögen entfernt: Einige Äusserungen enthalten keine dialektspezifischen Hinweise und können daher auch von Experten nicht zuverlässig klassifiziert werden. Hier könnte die Hinzufügung von akustischen Daten die fehlende Detailgenauigkeit der Transkriptionen ausgleichen.

Die Anfälligkeit für Transkriptoreneffekte, welche höher als erwartet ausfiel, veranlasste uns, die Transkriptionen einiger Texte zu überprüfen und systematische Transkriptionsunterschiede

zu beseitigen, was zu einigen Abweichungen zwischen den (nicht korrigierten) Texten der ersten Veröffentlichung und den (korrigierten) Texten der zweiten Veröffentlichung führte.

4.1.2 Dialektklassifikation

In dieser zweiten Fallstudie möchten wir das Problem der Dialekterkennung und -klassi-fizierung ausdehnen, indem wir a) alle Dokumente des ArchiMob-Korpus berücksichtigen und uns b) nicht auf vordefinierte Dialektgebiete stützen. Konkret wollen wir untersuchen, inwieweit Dialektgebiete direkt aus den Daten hergeleitet werden können.

Dialektgebiete aufzudecken ist eines der Hauptziele der Dialektometrie (cf. z. B. Goebl 1982). Die traditionelle dialektometrische Pipeline besteht aus den folgenden Schritten (nach Goebl 2010: 439):

- Die linguistischen Daten, typischerweise aus einem dialektologischen Atlas entnommen, werden in einer *Datenmatrix* aus n Ortspunkten × m linguistischen Merkmalen aufbereitet. Jede Zelle enthält die lokale Variante eines Merkmals an einem Ortspunkt.
- Aus der Datenmatrix wird durch paarweisen Vergleich der Merkmalsvektoren zweier Ortspunkte eine *Distanzmatrix* von n Punkten × n Punkten abgeleitet. Die Distanzmatrix ist typischerweise symmetrisch, mit Nullwerten auf der Diagonalen.
- Dann wird ein Dimensionsreduktions-Algorithmus angewendet, um jede Zeile der Distanzmatrix auf einen einzelnen Wert (oder eine kleine Anzahl von Werten v) zu reduzieren, was zu einer *Wertematrix* von n Punkten × v Werten führt. Dafür sind zahlreiche Algorithmen vorgeschlagen worden. Einer der einfachsten (und der nachfolgend verwendete) ist die hierarchische Clusteranalyse, bei der jedem Ortspunkt eine Clusternummer zugewiesen wird, so dass die Punkte mit den ähnlichsten linguistischen Merkmalen im gleichen Cluster zu liegen kommen.
- Die Werte der Wertematrix werden je nach Clusternummer farblich unterschiedlich auf einer Karte dargestellt. Die Hypothese ist, dass geografisch nahe beieinanderliegende Orte auch linguistisch nahe beieinander (i. e. im selben Cluster) liegen. Wo dies nicht der Fall ist, muss eine dialektologische Erklärung für diese Diskrepanz gefunden werden.

Die dialektometrische Pipeline wurde hauptsächlich auf Atlasdaten angewendet (cf. Goebl 2005), wobei typischerweise jede Spalte in der Datenmatrix eine linguistische Variable enthält. Im Gegensatz dazu steht die korpusbasierte Dialektometrie (cf. z. B. Wolk/Szmrecsanyi 2016), die anstelle von Atlasdaten Fliesstext aus Korpora zur Grundlage nimmt. Ein Vorteil der korpusbasierten Dialektometrie ist die Verfügbarkeit von Häufigkeitsinformationen, es gibt aber auch ein paar gewichtige Nachteile wie die ungleichmässige räumliche Abdeckung (natürlich vorkommende Texte neigen dazu, sich in bestimmten Regionen zu bündeln) und die geringe Dichte an linguistischen Phänomenen (die interessanten Merkmale zeigen sich typischerweise nur selten im Text). Beispielsweise gibt es nur gerade 114 normalisierte Worttypen, die in allen 43 ArchiMob-Dokumenten vorkommen. Darüber hinaus ist ein grosser Anteil der Variation auf persönliche Präferenzen und auf den Kontext zurückzuführen, und nicht in erster Linie auf den Dialekt des Sprechenden. Die Untersuchung linguistischer Distanzen mit Hilfe von Korpusdaten erfordert daher eine Abkehr von einer typischen Dialektdatenmatrix.

In einer Pilotstudie (cf. Scherrer 2012) haben wir auf der Grundlage einer vorläufigen Version des ArchiMob-Korpus versucht, Wörter mit ähnlichen Transkriptionen über Dialekte hinweg explizit zu paaren. Hier schlagen wir die Verwendung von Sprachmodellierung vor – eine Technik, die auch von Teilnehmern der Dialekterkennungs-Aufgabe verwendet wurde (cf. Gamallo et al. 2017) –, um direkt eine Distanzmatrix zu erstellen. Die zweite Hälfte der dialektometrischen Pipeline kann dann wie zuvor befolgt werden.

Wir erstellen für jedes Dokument des ArchiMob-Korpus ein separates Sprachmodell und berechnen, wie gut es zu allen anderen Dokumenten des Korpus passt, unter der Hypothese, dass ein Sprachmodell besser zu einem Text desselben Dialekts passt als zu einem Text eines entfernteren Dialekts. Dieses Mass des „Passens“ wird als Perplexität bezeichnet. Wir trainieren zeichenbasierte 4-Gramm-Sprachmodelle mit KenLM (cf. Heafield 2011) und berechnen die Perplexitäten auf Dokumentenebene. Die resultierende „Distanzmatrix“ ist nicht symmetrisch, da die Perplexität von Modell A auf Text B nicht garantiert mit der Perplexität von Modell B auf Text A identisch ist. Ebenso enthält die Diagonale nicht unbedingt Nullwerte, da die Perplexität von Modell A auf Text A nicht immer gleich 0 ist. Für die Klassifizierung verwenden wir die hierarchische Clusteranalyse mit dem Ward-Algorithmus.

Abbildung 5 zeigt die Ergebnisse mit 10 Clustern. Die oben erwähnten Transkriptoreneffekte kommen wieder recht stark zur Geltung: alle von der Transkriptorin AZ bearbeiteten Dokumente gehören zum dunkelgrauen Cluster, unabhängig von ihrer geographischen Verortung. Im Gegensatz dazu bilden die Dokumente von EP sechs Cluster, die alle geografisch homogen sind. Die Dokumente von Transkriptor PM landen im selben Cluster wie die von EP für den Raum Zürich (hellgrün), aber in einem eigenständigen Cluster für den Raum Bern (orange vs. gelb).

Diese Clusteranalyse kann mit einem ähnlichen Experiment verglichen werden (Hintergrundfarben in Abbildung 5), das auf Daten aus zwei schweizerdeutschen Dialektatlanten nach dem traditionellen dialektometrischen Ansatz beruht (cf. Scherrer/Stoeckle 2016). Die gute Übereinstimmung der Dokumente der Transkribierenden EP und PM deutet darauf hin, dass zumindest diese Untergruppe von Dokumenten ein dialektologisch differenziertes Signal enthält, das mit wesentlich kostenintensiveren Atlasdaten konkurrieren kann.

Da Transkriptoreneffekte nicht vollständig vermieden werden können (teilweise auch wegen der Änderungen der Transkriptionsleitlinien), werden sich zukünftige Aktivitäten auf die statistische Modellierung der Transkriptorenvariation und der dialektalen Variation als eigenständige Effekte konzentrieren (cf. Wieling et al. 2011).

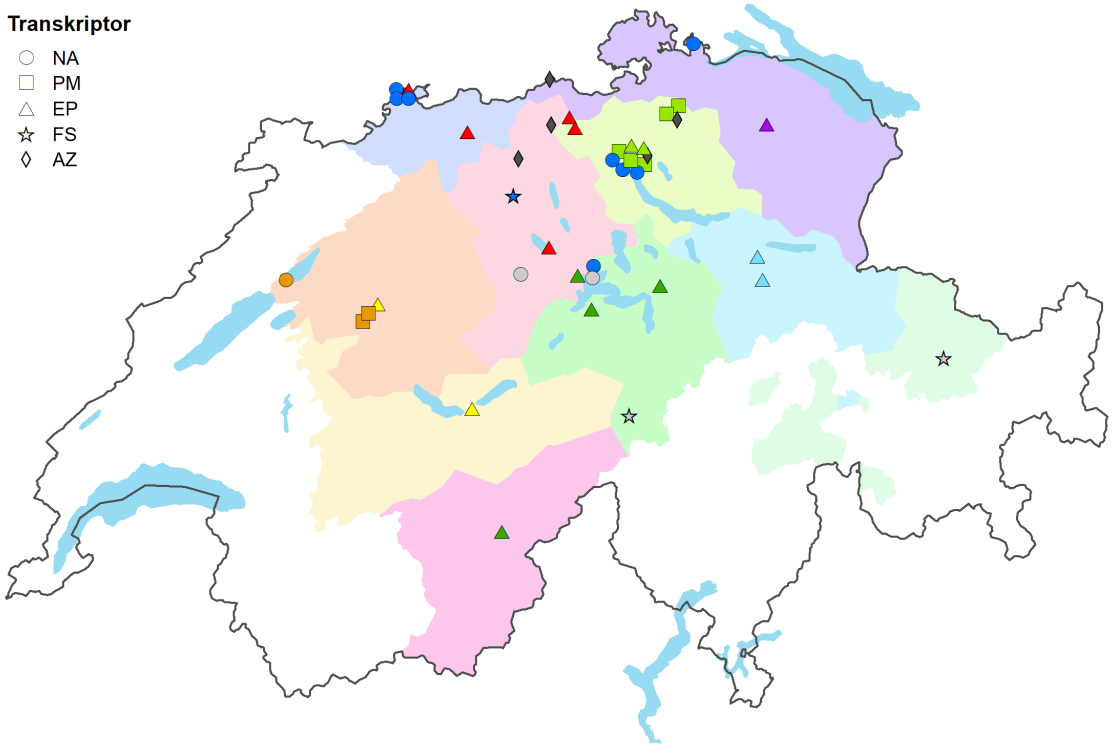


Abbildung 5: Clusteranalyse nach Ward basierend auf gegenseitigen Sprachmodell-Perplexitäten. Jedes Symbol steht für einen ArchiMob-Text; verschiedene Symbolformen stehen für unterschiedliche Transkribierende, verschiedene Farben für unterschiedliche Cluster. Die Hintergrundfarben entsprechen den mit Atlasdaten berechneten Clustern (Scherrer und Stoeckle 2016, Abbildung 2); die Farben werden manuell zugeordnet.

4.2 Die Normalisierungsebene als Grundlage für dialektologische Vergleiche

Im vorhergehenden Abschnitt wurde auf die Schwierigkeit des Vergleichs von Merkmalen in den verschiedenen Dialekttexten hingewiesen: nicht alle Gewährspersonen verwenden die gleichen linguistischen Strukturen und das gleiche Vokabular, und es ist schwierig, echte dialektologische Effekte von subjektiven und persönlichen Präferenzen zu unterscheiden. Dabei kann jedoch die Normalisierungsebene helfen: wir gehen davon aus, dass alle linguistischen Elemente (Wörter, Grapheme oder Zeichen), die auf gleiche Weise normalisiert sind, miteinander vergleichbar sind. In diesem Abschnitt stellen wir zwei Anwendungen dieses Ansatzes vor.

4.2.1 Phonologische Variationsmuster in Dialekttexten

Phonologische Eigenschaften in transkribierter Rede zu untersuchen ist eine Herausforderung, vor allem wenn bekannt ist, dass sich die Transkriptionsleitlinien im Laufe der Zeit verändert haben und Transkriptorenunterschiede bestehen. Trotz dieser Herausforderungen zeigen wir im Folgenden, dass bekannte phonologische Variationsmuster dank der Normalisierungsebene effizient extrahiert, verglichen und dargestellt werden können.

Wir definieren – unter Verwendung mehrerer methodologischer Abkürzungen – eine phonologische Variable als ein Graphem auf der Normalisierungsebene und ihre möglichen dialektalen Realisierungen (Varianten) als die Menge der Grapheme auf Transkriptionsebene, die mit dem

normalisierten Graphem aligniert sind. Zum Beispiel besteht die phonologische Variable, die durch das Normalisierungsgraphem *ck* vertreten wird, aus den dialektalen Varianten *k* und *gg*, deren Häufigkeitsverteilung je nach Herkunft der Texte variiert.[6] Damit diese Definition funktioniert, müssen wir a) die Zeichen der transkribierten und normalisierten Wörter miteinander alignieren und b) benachbarte Zeichen bei Bedarf in mehrstellige Grapheme gruppieren.

Eine beliebte Technik der Zeichenalignierung basiert auf der Levenshtein-Distanz, bei der die Bearbeitungsschritte, die zur Berechnung der Levenshtein-Distanz beitragen, in Alignmentpaare umgewandelt werden. Die zeichenbasierte statistische maschinelle Übersetzung (CSMT), welche wir bereits für die Normalisierung verwendet haben, stellt eine Alternative zu diesem Ansatz dar. Sie setzt keine Wort- (oder Zeichen-) Identität voraus und funktioniert daher auch mit unterschiedlichen Zeicheninventaren oder Schriftsystemen. Die am weitesten verbreiteten Alignierungsmodelle wurden in den Anfängen der statistischen maschinellen Übersetzung eingeführt (cf. Brown et al. 1993) und wurden für die Zeichenalignierung in unserem CSMT-Verfahren verwendet. Da die Zeichenalignierung ein integraler Bestandteil der CSMT-Übersetzungsmodelle ist, können wir aus unseren Normalisierungsmodellen alle interessanten Alignments entnehmen.

Grapheme bestehen nicht immer aus einzelnen Zeichen. Häufig auftretende und häufig alignierte Zeichenfolgen sollten zu einem mehrstelligen Graphem zusammengefasst werden. Dieser Prozess wurde auch im Bereich der statistischen maschinellen Übersetzung unter dem Namen Phrasenextraktion untersucht (cf. Och et al. 1999) und kann wieder einfach von der Wort- auf die Zeichenebene übertragen werden. Die während des CSMT-Modelltrainings erstellte Phrasentabelle listet alle Graphempaare mit ihren (Ko-)Okkurrenz-Zahlen auf, so dass sich die relativen Häufigkeiten der Transkriptionsgrapheme einfach berechnen lassen.

Wir kommen auf das obige Beispiel zurück und untersuchen, wie das normalisierte Graphem *ck* in zwei beliebig gewählten ArchiMob-Dokumenten ausgedrückt wird:

- Dokument 1: *k* 37.0%, *gg* 63.0%
- Dokument 2: *k* 95.2%, *gg* 2.4%, *ch* 2.4%.

Diese Analyse kann auf alle Dokumente des ArchiMob-Korpus ausgedehnt werden und die Häufigkeitsverteilung jeder Variante auf einer Karte dargestellt werden. Abbildung 6 zeigt eine solche Darstellung für die *gg*-Variante. Die aus den ArchiMob-Texten entnommenen Häufigkeiten können mit den Atlasdaten aus dem *Sprachatlas der deutschen Schweiz* verglichen werden (SDS; cf. Hotzenköcherle et al. 1962–1997). Das Verbreitungsgebiet der *gg*-Variante nach dem Atlas (Karte 2/095) ist in Abbildung 6 grün hinterlegt. Sieben ArchiMob-Dokumente zeigen relative Häufigkeiten über 50% für die *gg*-Variante. Alle diese Dokumente liegen in den drei Regionen, in denen auch der SDS die *gg*-Variante zeigen: Basel (Nordwesten), St. Gallen (Nordosten) und Glarus (Südosten).

[6] Nach dem Leitfaden von Dieth widerspiegelt das Graphem ‹*k*› die Aussprache [kx], während das Graphem ‹*gg*› [k:] widerspiegelt. Die phonetische Realisierung der Normalisierungsgrapheme ist hier nicht relevant.

Ein weiteres interessantes Phänomen ist die Vokalisierung des intervokalischen *ll* in westschweizerdeutschen Dialekten. Abbildung 7 zeigt die relativen Häufigkeiten der vokalen Variante *u* in den ArchiMob-Texten und in Atlasdaten. Auch hier ist zu erkennen, dass sich alle vokalisierenden ArchiMob-Gewährspersonen in den Gebieten (oder in deren Nähe) befinden, in denen der SDS Vokalisierung vorsieht. Die Häufigkeitswerte sind jedoch breit gestreut. Bei den drei Texten der Region Bern widerspiegelt diese Variation – zumindest teilweise – den soziolinguistischen Stellenwert der *l*-Vokalisierung als Phänomen der Unterschicht (cf. Siebenhaar 2000): Der Gärtner verwendet die Vokalisierung häufiger (69%) als die zwei Sprecher der oberen Mittelschicht, ein Konstrukteur (33%) und ein Arzt (49%). In der Zentralschweiz weisen zwei Dokumente Vokalisierung auf, und beide stammen von den Grenzen des Vokalisierungsgebietes, wie es in der SDS-Karte definiert ist. Umgekehrt dazu gibt es auch einige ArchiMob-Dokumente aus SDS-Vokalisierungsgebieten, die aber keine Nachweise für dieses Phänomen liefern. Ob diese Diskrepanz auf Sprachwandel oder auf Transkriptoreneffekte zurückzuführen ist, muss noch analysiert werden.

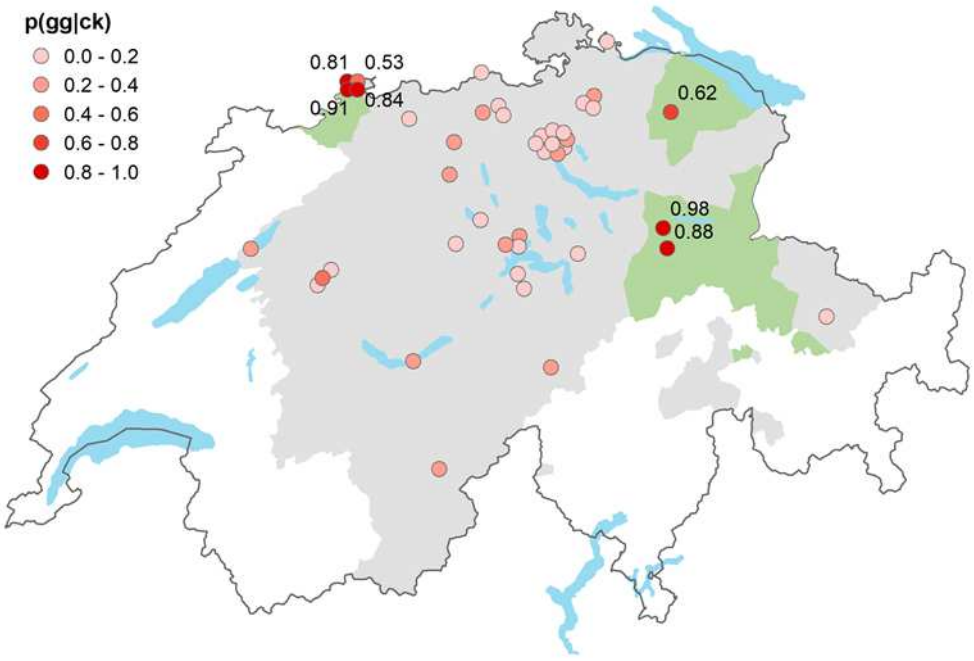


Abbildung 6: Wahrscheinlichkeiten von dialektalem *gg* entsprechend normalisiertem *ck*. Die grünen Flächen stellen die Verteilung der *gg*-Variante in der SDS-Karte 2/095 'drücken' dar.

Die beiden oben genannten Beispiele zeigen, dass die geographische Ausbreitung einiger phonologischer Varianten aus dem ArchiMob-Korpus auffallend gut mit denen des SDS übereinstimmt. Da die ArchiMob-Informanten etwa eine Generation jünger sind als die SDS-Gewährspersonen, kann die vorgeschlagene Technik auch verwendet werden, um Dialektwandel zu untersuchen. So schreibt Christen (2001), dass sich die *l*-Vokalisierung nach Osten bis zur Stadt Luzern und Nidwalden (Südufer des Vierwaldstättersees) ausgebreitet hat, die ArchiMob-Dokumente aus dieser Gegend zeigen jedoch (noch) keine Vokalisierung. Dies deutet darauf hin, dass dieser sprachliche Wandel in jüngerer Zeit begonnen haben könnte.

Es muss allerdings angemerkt werden, dass die hier vorgestellte Methode durch die Genauigkeit der Transkription (und die Korrektheit der Normalisierung) begrenzt ist. Es können natürlich nur Variationsmuster ermittelt werden, die sich in der Transkription widerspiegeln. Studien zur Realisierung von /r/ können beispielsweise nicht durchgeführt werden (zumindest nicht ohne Analyse der entsprechenden Audiodaten), da die verschiedenen Varianten in der Transkription nicht unterschieden werden. Ebenso wären Studien zu Vokaleigenschaften nicht verlässlich, da nicht alle Dokumente des ArchiMob-Korpus auf gleiche Weise transkribiert wurden. Dennoch kann der vorgeschlagene Ansatz Aufschluss über gewisse Aspekte von Dialektvariation und -wandel in der Deutschschweiz geben.

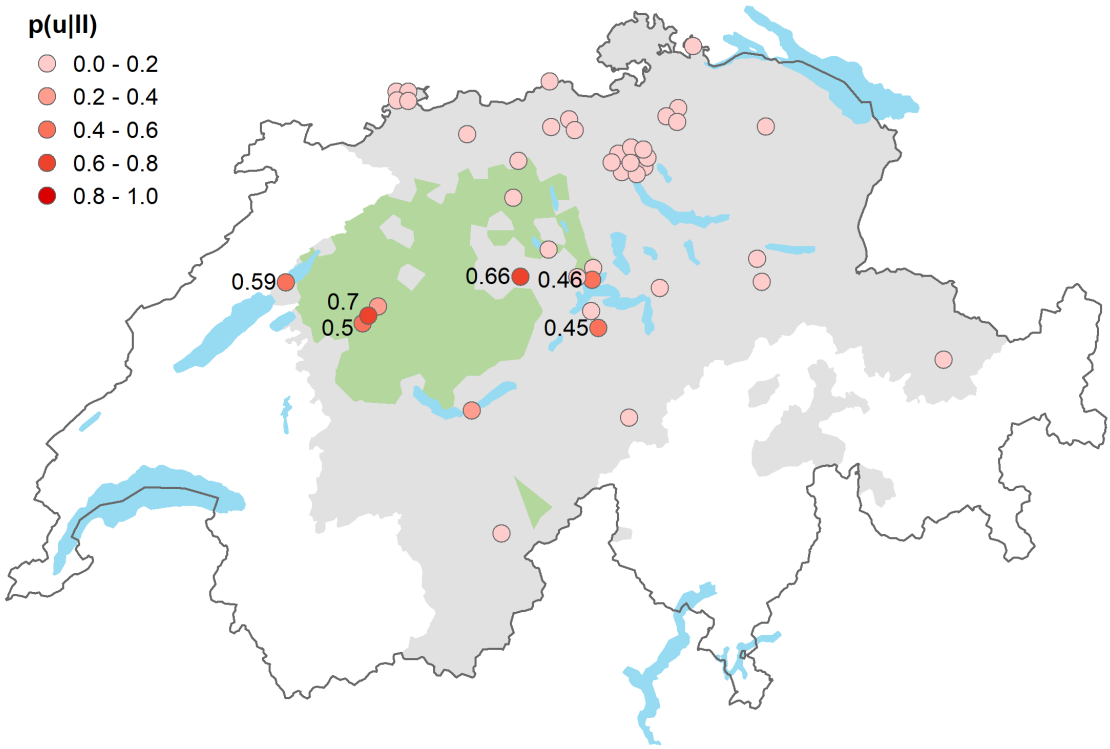


Abbildung 7: Wahrscheinlichkeiten von dialektalem *u* entsprechend normalisiertem *ll*. Die grünen Flächen stellen die Verteilung der *u*-Variante in der SDS-Karte 2/198 'Teller' dar.

4.2.2 Dialektalitätsmessungen

Die Normalisierungsebene kann auch für Gesamtanalysen der ArchiMob-Daten verwendet werden. Dafür bauen wir auf dem Konzept der Dialektalitätsmessung auf (cf. Herrgen/Schmidt 1989; Herrgen et al. 2001). *Dialektalität* drückt die Entfernung eines Dialekttextes zur Standardvarietät aus und hat sich z. B. als relevantes Mass zur Einschätzung von Dialektabbau erwiesen.

Die Dialektalitätsmessung erfordert phonetisch transkribierte, auf Zeichenebene alignierte Daten im Dialekt und in der Standardsprache. Die Phoneme werden paarweise verglichen, und für jedes Phonempaar wird ein Entfernungswert berechnet, basierend auf der Anzahl der phonetischen Merkmale, die zu ändern sind. Diese Entfernungswerte werden dann über Wörter und Äusserungen gemittelt, um einen einzigen Dialektalitätswert pro Text zu erhalten.

Wir vereinfachen diese Idee massiv für unsere Zwecke. Erstens gehen wir davon aus, dass die Normalisierungsebene unsere Standardsprache ist, was nicht ganz korrekt ist. Zweitens versuchen wir nicht, die Transkriptionen und Normalisierungen in richtige phonetische Transkriptionen umzuwandeln, da sie in der Regel unterspezifiziert sind. Stattdessen verwenden wir die einfache Levenshtein-Distanz, um einen Entfernungswert pro Wort zu berechnen. Abbildung 8 stellt die Dialektalitätswerte aller ArchiMob-Texte gemäss dieser vereinfachten Berechnung dar.

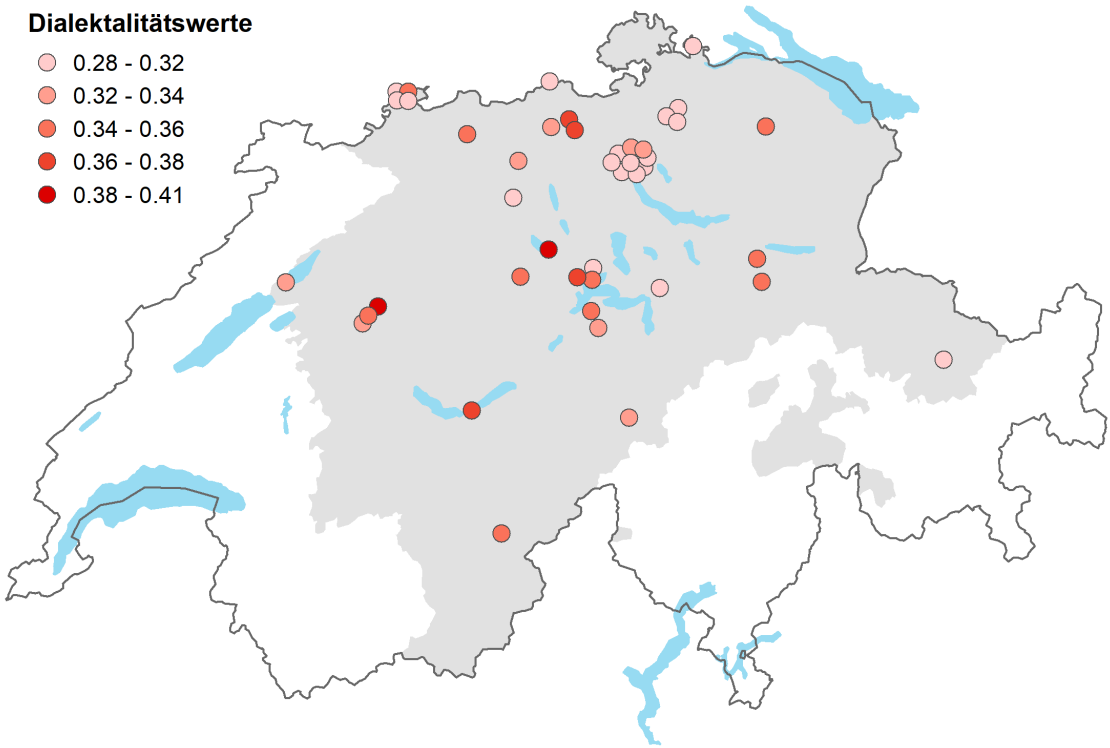


Abbildung 8: Dialektalitätswerte (normalisierte und gemittelte Levenshtein-Distanzwerte)

Die Ergebnisse zeigen, dass sich die Dialektalitätswerte zwischen den Dokumenten nicht wesentlich unterscheiden. Auch scheinen sie weder mit der Geographie noch mit dem / der Transkribierenden stark zu korrelieren. Es wird jedenfalls deutlich, dass die Dialekte in der Nordschweiz keinem höheren Assimilationsdruck durch das (benachbarte) Standarddeutsche ausgesetzt sind als die südschweizerischen Dialekte. Dies ist zu erwarten, da die Standardvarietät in der Schweiz als vertikales Pendant in der Diglossie und nicht als „horizontales“ Pendant auf der geografischen (im vorliegenden Fall) Nord-Süd-Achse wahrgenommen wird (cf. Siebenhaar/Wyler 1997).

Die niedrigsten Dialektalitätswerte finden sich im Raum Zürich-Winterthur, was darauf hindeutet, dass der Zürcher Dialekt als eine Art Default-Dialekt mit einer geringen Anzahl charakteristischer Merkmale angesehen werden kann. Dieser Effekt wurde in mehreren Studien auf der Grundlage verschiedener Datensätze nachgewiesen (cf. Scherrer/Rambow 2010; Hollenstein/Aepli 2015; Scherrer/Stoeckle 2016).

4.3 Inhaltsanalyse

In den vorangegangenen Abschnitten haben wir uns auf die formelle Variation in den ArchiMob-Texten konzentriert. Eine andere, über die Grenzen der Dialektologie hinausgehende Analyseebene betrifft den Inhalt der Texte. Nachfolgend wird ein mögliches Analyseszenario skizziert.

Die *Kollokationsanalyse* (oder allgemeiner die Kookkurrenzanalyse) ist eine simple und gängige Methode, um Verbindungen zwischen sprachlichen Ausdrücken in Texten zu messen. Für jedes Schlüsselwort werden die am häufigsten kookkurrierenden Wörter mit Hilfe von Gewichtungstechniken aus dem Information Retrieval gesucht. Die Kollokationsanalyse kann für ein ganzes Korpus oder separat für verschiedene Teile des Korpus durchgeführt werden, um Veränderungen im Verwendungsmuster zu ergründen.

Das Sketch Engine-Korpusanalysewerkzeug bietet eine integrierte Kollokationsanalysefunktion, die von Kontoinhabern interaktiv genutzt werden kann. Als Beispiel zeigt Abbildung 9 eine Kollokationsanalyse für das Wort *chind* ‚Kind/Kinder', die über alle Dokumente des ArchiMob-Korpus hinweg durchgeführt wird.[7] Die Kollokate werden in einem Kreis um das Schlüsselwort herum dargestellt, wobei verschiedene Farben verschiedene grammatikalische Kategorien darstellen.

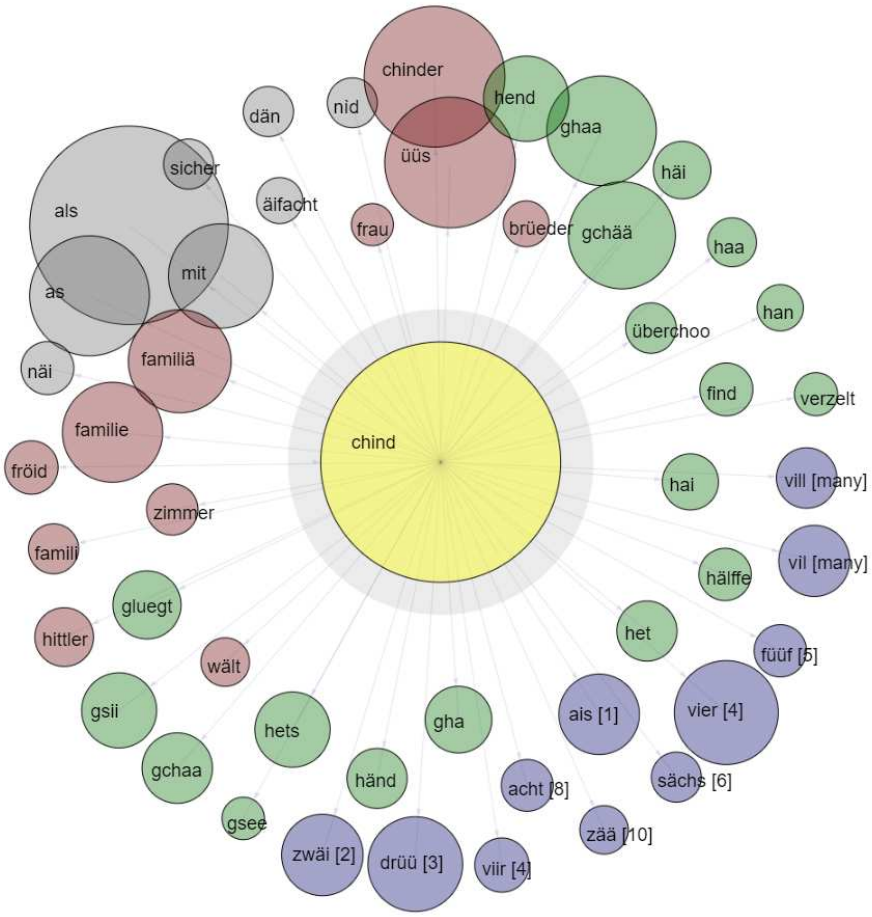


Abbildung 9: Kollokate des Wortes *chind* im ArchiMob-Korpus.

[7] Zu *chind*: Die Singular- und Pluralformen sind in vielen schweizerdeutschen Dialekten homophon.

Allein durch die Betrachtung der Zahlwörter (blaue Kreise) kann uns das ArchiMob-Korpus einen interessanten Einblick in die traditionellen Familiengrössen der Schweiz in der ersten Hälfte des 20. Jahrhunderts geben. Wörter, die sich auf andere Konzepte und deren Assoziationen beziehen, können auf ähnliche Weise analysiert werden.

In einer ähnlich gefassten Studie demonstriert Schifferle (2017) überzeugend den Nutzen des ArchiMob-Korpus für lexikologische Analysen, indem er die Verwendung der Beziehungsbegriffe *koleeg* ‚Kollege' und *fründ* ‚Freund' in den 16 ArchiMob-Texten der 1. Phase untersucht. Dabei fällt auf, dass *koleeg* fast ausschliesslich von männlichen Informanten und *fründ* fast ausschliesslich von weiblichen Gewährspersonen verwendet wird. Auch sind (schwache) Indizien eines Bedeutungswandels des Begriffs *koleeg* von ‚Arbeits-/Vereins-/Militärkollege' zu einem allgemeineren Begriff enger persönlicher Freundschaft sichtbar.

Zuletzt möchten wir auch darauf hinweisen, dass sich das ArchiMob-Korpus auch für qualitative Forschungsansätze eignet, beispielsweise mittels Methoden der interaktionalen Soziolinguistik oder der Konversationsanalyse. Dabei könnten Themenbereiche wie Spracheinstellungen und Identitäten untersucht werden. Die vielfältigen Zugriffsmöglichkeiten der oben vorgestellten Korpusabfragesysteme unterstützen solche Forschungsansätze.

5 Zusammenfassung

In diesem Artikel stellen wir den Aufbau und die potentiellen Anwendungsgebiete einer universellen Forschungsressource vor, die aus manuellen Transkriptionen von Zeitzeugenbefragungen auf Schweizerdeutsch besteht. Wir argumentieren, dass eine solche Ressource eine wichtige Grundlage für neue quantitative Ansätze zur Untersuchung von Sprachgebrauch und -variation ist, vor allem in der Dialektologie, aber auch in den Sozialwissenschaften und der Geschichte. Mit unseren Forschungsarbeiten zur Kodierung und Annotation von gesprochenen Schweizerdeutsch hat sich ein Know-how angesammelt, das gewinnbringend für die Entwicklung ähnlich gelagerter Ressourcen eingesetzt werden kann. Ebenso steht mit dem ArchiMob-Korpus ein Endprodukt für verschiedenste Arten korpusbasierter Forschung zur Verfügung.

Literaturverzeichnis

Archimob Projekt (2000-2008): http://archimob.ch [29.07.2019]

ArchiMob Korpus (2016-2019): https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html [29.07.2019]

Baron, Alistair/Rayson, Paul (2008): "VARD 2: A tool for dealing with spelling variation in historical corpora". In: *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham, UK, Aston University.

Brown, Peter F. et al. (1993): "The mathematics of statistical machine translation: Parameter estimation". *Computational linguistics* 19/2: 263–311.

Christen, Helen (1998): *Dialekt im Alltag: eine empirische Untersuchung zur lokalen Komponente heutiger schweizerdeutscher Varietäten*. Tübingen: Niemeyer.

Christen, Helen (2001): „Ein Dialektmarker auf Erfolgskurs: Die /l/-Vokalisierung in der deutschsprachigen Schweiz". *Zeitschrift für Dialektologie und Linguistik* 68/1: 16–26.

Christen, Helen/Glaser, Elvira/Friedli, Mathias (2013): *Kleiner Sprachatlas der deutschen Schweiz*. Frauenfeld: Huber.

De Clercq, Orphée et al. (2013): "Normalization of Dutch user-generated content". *Proceedings of RANLP 2013*, Hissar: 179–188.

Dieth, Eugen (1986): *Schwyzertütschi Dialäktschrift*. 2. Ausgabe. Aarau: Sauerländer.

Friedli, Mathias (2012): *Der Komparativanschluss im Schweizerdeutschen: Arealität, Variation und Wandel*. Dissertation, Universität Zürich. https://doi.org/10.5167/uzh-68746 [29.07.2019]

Gamallo, Pablo/Pichel, Jose Ramom/Alegria, Iñaki (2017): "A perplexity-based method for similar languages discrimination". *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, ACL: 109–114.

Glaser, Elvira/Bart, Gabriela (2015): „Dialektsyntax des Schweizerdeutschen". In: Kehrein, Roland/Lameli, Alfred/Rabanus, Stefan (eds.*): Regionale Variation des Deutschen. Projekte und Perspektiven*. Berlin/New York, De Gruyter: 79–105.

Goebl, Hans (1982): *Dialektometrie. Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Goebl, Hans (2005): „Dialektometrie" (Art. 37). In: Köhler, Reinhard/Altmann, Gabriel/Piotrowski Rajmund G. (eds.): *Quantitative Linguistik/Quantitative Linguistics. Ein internationales Handbuch/An International Handbook*. Handbücher zur Sprach- und Kommunikationswissenschaft 27, Berlin/New York, De Gruyter: 498–531.

Goebl, Hans (2010): "Dialectometry and quantitative mapping". In: Lameli, Alfred/Kehrein, Roland/Rabanus, Stefan (eds.): *Language and Space. An International Handbook of Linguistic Variation*. Handbücher zur Sprach- und Kommunikationswissenschaft 30.2, Berlin/New York, De Gruyter: 433–457.

Heafield, Kenneth (2011): "KenLM: faster and smaller language model queries". *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*. Edinburgh, ACL: 187–197.

Herrgen, Joachim/Schmidt, Jürgen Erich (1989): „Dialektalitätsareale und Dialektabbau". In: Putschke, Wolfgang/Veith, Werner/Wiesinger, Peter (eds.): *Dialektgeographie und Dialektologie. Günter Bellmann zum 60. Geburtstag von seinen Schülern und Freunden*. Marburg, Elwert: 304–346. (= Deutsche Dialektgeographie 90).

Herrgen, Joachim et al. (2001): *Dialektalität als phonetische Distanz. Ein Verfahren zur Messung standarddivergenter Sprechformen*. http://archiv.ub.uni-marburg.de/es/2008/0007/pdf/dialektalitaetsmessung.pdf [29.07.2019]

Hinrichs, Erhard W. et al. (2000): "The Tübingen treebanks for spoken German, English, and Japanese". In: Wahlster, Wolfgang (ed.): *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin/Heidelberg, Springer: 550–574.

Hollenstein, Nora/Aepli, Noëmi (2014): "Compilation of a Swiss German dialect corpus and its application to PoS tagging". *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, COLING 2014. Dublin, ACL: 85–94.

Hollenstein, Nora/Aepli, Noëmi (2015): "A resource for natural language processing of Swiss German dialects". *Proceedings of GSCL 2015*. Duisburg-Essen, GSCL: 108–109.

Honnet, Pierre-Edouard et al. (2018): "Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German". *Proceedings of LREC 2018*. Miyazaki, ELRA: 3781–3788.

Hotzenköcherle, Rudolf et al. (eds.) (1962–1997): *Sprachatlas der deutschen Schweiz*. Bern: Francke.

Kilgarriff, Adam et al. (2014): "The Sketch Engine: ten years on." *Lexicography* 1/1: 7–36.

Kisler, Thomas/Schiel, Florian/Sloetjes, Han (2012): "Signal processing via web services: the use case WebMAUS". *Proceedings of Digital Humanities Conference 2012*. Hamburg: 30–34.

Kolde, Gottfried (1981): *Sprachkontakte in gemischtsprachigen Städten. Vergleichende Untersuchungen über Voraussetzungen und Formen sprachlicher Interaktion verschiedensprachiger Jugendlicher in den Schweizer Städten Biel/Bienne und Fribourg/Freiburg i.Ue*. Wiesbaden: Steiner. (= Zeitschrift für Dialektologie und Linguistik, Beihefte 37).

Krause, Thomas/Zeldes, Amir (2016): "ANNIS3: A new architecture for generic corpus query and visualization". *Literary and Linguistic Computing* 31/1: 118–139.

Leemann, Adrian et al. (2016): "Crowdsourcing language change with smartphone applications". *PLoS ONE* 11/1.

Ljubešić, Nikola/Erjavec, Tomaž/Fišer, Daria (2014): "Standardizing tweets with character-level machine translation". *Proceedings of CICLing 2014, Lecture notes in computer science*. Kathmandu, Springer: 164–175.

Ljubešić, Nikola et al. (2016): "Normalising Slovene data: historical texts vs. user-generated content". *Proceedings of KONVENS 2016*. Bochum, Bochumer Linguistische Arbeitsberichte: 146–155.

Lusetti, Massimo et al. (2019): "Encoder-Decoder Methods for Text Normalization". *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Santa Fe, COLING: 18–28.

Och, Franz Josef/Tillmann, Christoph/Ney, Hermann (1999): "Improved alignment models for statistical machine translation". *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. College Park, ACL: 20–28.

Pettersson, Eva/Megyesi, Beáta B./Tiedemann, Jörg (2013): "An SMT approach to automatic annotation of historical text". *Proceedings of the Nodalida Workshop on Computational Historical Linguistics*. Oslo, Linköping University Electronic Press: 54–69.

Pettersson, Eva/Megyesi, Beáta B./Nivre, Joakim (2014): "A multilingual evaluation of three spelling normalisation methods for historical text". *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Göteborg, ACL: 32–41.

Richner-Steiner, Janine (2011): *‚E ganz e liebi Frau'. Zu den Stellungsvarianten des indefiniten Artikels in der adverbiell erweiterten Nominalphrase im Schweizerdeutschen. Eine dialektologische Untersuchung mit quantitativgeographischem Fokus*. Dissertation, Universität Zürich. https://opac.nebis.ch/ediss/20121398.pdf [29.07.2019]

Ruef, Beni/Ueberwasser, Simone (2013): "The taming of a dialect: Interlinear glossing of Swiss German text messages". In: Zampieri, Marcos/Diwersy, Sascha (eds.): *Non-standard Data Sources in Corpus-based Research*. Aachen, Shaker: 61–68.

Rychlý, Pavel (2007): "Manatee/Bonito – a modular corpus manager". *Proceedings of the First Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno, Masaryk University: 65–70.

Samardžić, Tanja/Scherrer, Yves/Glaser, Elvira (2015): "Normalising orthographic and dialectal variants for the automatic processing of Swiss German". *Proceedings of the 7th Language and Technology Conference*. Poznan, ELRA: 294–298.

Samardžić, Tanja/Scherrer, Yves/Glaser, Elvira (2016): "ArchiMob – a corpus of spoken Swiss German". *Proceedings of LREC 2016*. Portorož, ELRA: 4061–4066.

Scherrer, Yves (2012): "Recovering dialect geography from an unaligned comparable corpus". In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Avignon, ACL: 63–71.

Scherrer, Yves/Erjavec, Tomaž (2016): "Modernising historical Slovene words". *Natural Language Engineering* 22/6: 881–905.

Scherrer, Yves/Ljubešić, Nikola (2016): "Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation". *Proceedings of KONVENS 2016*. Bochum, Bochumer Linguistische Arbeitsberichte: 248–255.

Scherrer, Yves/Rambow, Owen (2010): "Word-based dialect identification with georeferenced rules". *Proceedings of EMNLP 2010*. Cambridge (MA), ACL: 1151–1161.

Scherrer, Yves/Stoeckle, Philipp (2016): "A quantitative approach to Swiss German – dialectometric analyses and comparisons of linguistic levels". *Dialectologia et Geolinguistica* 24/1: 92–125.

Scherrer, Yves/Samardžić, Tanja/Glaser, Elvira (erscheint): "Digitising Swiss German – How to process and study a polycentric spoken language". *Language Resources and Evaluation*.

Schifferle, Hans-Peter (2017): „Helvetische Beziehungen? Gschpäändli, Koleege, Fründ. Beziehungsbezeichnungen im Schweizerdeutschen". In: Linke, Angelika/Schröter, Juliane (eds.): *Sprache und Beziehung*. Berlin/Boston, De Gruyter: 183–206. (= Linguistik – Impulse & Tendenzen 69).

Schmidt, Thomas (2012): "EXMARaLDA and the FOLK tools." *Proceedings of LREC 2012*. Istanbul, ELRA: 236–240.

Schönenberger, Manuela/Haeberli, Eric (erscheint): „Ein geparstes und grammatisch annotiertes Korpus schweizerdeutscher Spontansprachdaten". *Zeitschrift für Germanistische Linguistik*.

Siebenhaar, Beat (2000): „Stadtberndeutsch – Sprachschichten einst und jetzt". In: Siebenhaar, Beat/Stäheli, Fredy (Hrsg.): *Stadtberndeutsch – Sprachporträts aus der Stadt Bern*. Murten, Licorne Verlag: 7–32. (= Schweizer Dialekte in Text und Ton 5.1).

Siebenhaar, Beat (2003): „Sprachgeographische Aspekte der Morphologie und Verschriftung in schweizerdeutschen Chats". *Linguistik online* 15. https://bop.unibe.ch/linguistik-online/issue/view/200 [29.07.2019]

Siebenhaar, Beat/Wyler, Alfred (1997): *Dialekt und Hochsprache in der deutschsprachigen Schweiz*. 5. Auflage. Zürich: Pro Helvetia.

Stark, Elisabeth/Ueberwasser, Simone/Ruef, Beni (2009–2015): *Swiss SMS corpus*. Universität Zürich. https://sms.linguistik.uzh.ch [29.07.2019]

Staub, Friedrich et al. (eds.) (1881ff.): *Schweizerisches Idiotikon. Wörterbuch der schweizerdeutschen Sprache*. Frauenfeld, Huber 1881–2012/Basel, Schwabe 2015ff.

Schiller, Anne et al. (1999): *Guidelines für das Tagging deutscher Textkorpora mit STTS*. Universität Stuttgart/Universität Tübingen.

Tiedemann, Jörg (2009): "Character-based PSMT for closely related languages". *Proceedings of EAMT 2009*. Barcelona, EAMT: 12–19.

Tjong Kim Sang, Erik et al. (2017): "The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation". *Computational Linguistics in the Netherlands Journal* 7: 53–64.

Vilar, David/Peter, Jan-T./Ney, Hermann (2007): "Can we translate letters?" *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, ACL: 33–39.

Wieling, Martijn/Nerbonne, John/Baayen, R. Harald (2011): "Quantitative social dialectology: Explaining linguistic variation geographically and socially". *PLoS ONE* 6/9.

Wolk, Christoph/Szmrecsanyi, Benedikt (2016): "Top-down and bottom-up advances in corpus-based dialectometry". Côté, Marie-Hélène/Knooihuizen, Remco/Nerbonne, John (eds.): *The future of dialects: Selected papers from Methods in Dialectology XV*. Berlin, Language Science Press: 225–244. (= Language Variation 1).

Zampieri, Marcos et al. (2014): "A report on the DSL shared task 2014". *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. Dublin, ACL: 58–67.

Zampieri, Marcos et al. (2017): "Findings of the VarDial evaluation campaign 2017". *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, ACL: 1–15.

Zampieri, Marcos et al. (2018): "Language identification and morphosyntactic tagging: The second VarDial evaluation campaign". *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Santa Fe, COLING: 1–17.

Anhang: Übersicht der Dokumente im ArchiMob-Korpus

Phase/ Transkriptor	ID	Herkunft Gewährsperson	Geschlecht	Anzahl Äusserungen
1 / EP	1007	Stans NW	w	972
	1048	Ennenda GL	w	1542
	1063	Baden AG	w	1561
	1073	Gelterkinden BL	w	1573
	1075	Basel BS	m	1056
	1142	Ittigen BE	m	798
	1143	Zürich-Seebach ZH	m	2060
	1147	Baden AG	m	2441
	1170	Matten b. Interlaken BE	w	935
	1195	Sempach LU	m	1102
	1198	Zuzwil SG/Brütten ZH	m	2438
	1207	Näfels GL	m	2012
	1209	Schwyz SZ	m	721
	1212	Naters VS	w	956
	1261	Luzern LU	w	1390
	1270	Wallisellen ZH	w	1875
2 / PM	1082	Zürich-Höngg ZH	w	1544
	1083	Winterthur ZH	w	1316
	1087	Zürich ZH	m	1050
	1121	Köniz/Bern BE	m	1082
	1215	Bangerten/Bern BE	m	2312
	1225	Zürich ZH	m	962
	1244	Winterthur-Wülflingen ZH	m	1540
3 / NA	1008	Meggen LU	m	1301
	1055	Zürich ZH	m	945
	1138	Meggen LU	w	599
	1188	Zürich ZH/Basel BS	m	1598
	1189	Zürich ZH	m	2419
	1205	Ramsen SH	m	1380
3 / AZ	1228	Zürich ZH	m	1706
	1248	Baden/Brugg AG	m	2541
	1259	Full-Reuenthal AG	m	2280
	1295	Buchs/Zofingen AG	w	1332
	1300	Winterthur ZH	m	3075
4 / NA	1044	Basel BS	m	2457
	1053	Wolfenschiessen NW	m	1219
	1203	Erlach BE	m	1921
	1224	Basel BS	m	2387
	1235	Luthern/Wolhusen LU	w	855
	1263	Basel BS	m	1615

4 / FS	1163	Staffelbach/Gränichen AG	m	2890
	1240	Davos/Chur GR	w	350
	1255	Göscheneralp UR	m	927